THE UNIVERSITY
of EDINBURGH

The predicted risk of harm versus treatment benefit in large randomised controlled trials

Douglas David Thompson

Doctor of Philosophy – The University of Edinburgh – 2015

Contents

Preface

I, Douglas David Thompson, declare that the following thesis has been composed by me and has not been submitted for any other degree or professional qualification.

Acknowledgements

I would like to thank my two supervisors, Professor Gordon Murray and Dr William Whiteley. Their expert guidance and supervision throughout my PhD has helped develop me as a statistician. Without their efforts this thesis would never have been possible.

I am extremely proud to have been part of the MRC Edinburgh Hub for Trials Methodology Research as well as the Centre for Population Health Sciences at the University of Edinburgh. I'd like to thank the many statisticians, epidemiologists and clinicians that I've met over this time who have been kind enough to chat to me about their work and my work: Dr Niall Anderson, Dr Fay Crawford, Dr Nynke Halbesma, Dr Christian Hansen, Peter Joshi, Ashma Krishan, Robert Lee, Dr Steff Lewis, Dr David McAllister, Dr Susannah McLean, Dr Craig Reed and Dr Chris Weir.

I am very grateful to the MRC for funding this research (MRC HTMR grant [G0800803]) and to those who made their data available for analysis: Professor Peter Sandercock (IST-1 and IST-3); Professor Zhengming Chen (CAST); Dr Livia Candelise (MAST-I); and Professor Cathie Sudlow (ESS).

I would also like to extend my gratitude to Professor Ewout Steyerberg and Dr Hester Lingsma for their warm welcome during a fascinating visit at Erasmus MC, Rotterdam. I learnt a lot from our work together.

To everyone in room 316: Margaret Horne, Stephanie Read, Hannah Ensor, Jacquie Stephen and Dr Kieren Egan (an honouree member and not bad squash player!). I couldn't have asked for a nicer group of folk to have worked alongside during the past three years. A big thank you also to Morag Leitch for tirelessly looking after us all!

Finally to Kirsten Stewart, for her support, encouragement and patience in letting me devote evenings and weekends to the completion of this thesis.

Abstract

Most drugs come with unwanted, and perhaps harmful, side-effects. Depending on the size of the treatment benefit such harms may be tolerable. In acute stroke, treatment with aspirin and treatment with alteplase have both proven to be effective in reducing the odds of death or dependency in follow-up. However, in both cases, treated patients are subject to a greater risk of haemorrhage – a serious side-effect which could result in early death or greater dependency. Current treatment licenses are restricted so as to avoid treating those with certain traits or risk factors associated with bleeding. It is plausible however that a weighted combination of all these factors would achieve better discrimination than an informal assessment of each individual risk factor. This has the potential to help target treatment to those most likely to benefit and avoid treating those at greater risk from harm. This thesis will therefore: (i) explore how predictions of harm and benefit are currently made; (ii) seek to make improvements by adopting more rigorous methodological approaches in model development; and (iii) investigate how the predicted risk of harm and treatment benefit could be used to strike an optimal balance.

Statistical prediction is not an exact science. Before clinical utility can be established it is essential that the performance of any prediction method be assessed at the point of application. A prediction method must attain certain desirable properties to be of any use, namely: good discrimination – which quantifies how well the prediction method can separate events from non-events; and good calibration – which measures how close the obtained predicted risks match the observed. A comparison of informal predictions made by clinicians and formal predictions made by clinical prediction models is presented using a prospective observational study of stroke patients seen at a single centre hospital in Edinburgh. These results suggest that both prediction methods achieve similar discrimination. A stratified framework based on predicted risks obtained from clinical prediction models is considered using data from large randomised trials. First, with three of the largest aspirin trials it is shown that there is no evidence to suggest that the benefit of aspirin on reducing six month death or dependency varies with the predicted risk of benefit or with the predicted risk of harm. Second, using data from the third International Stroke Trial (IST3) a similar

question is posed of the effect of alteplase and the predicted risk of symptomatic intracranial haemorrhage. It was found that this relationship corresponded strongly with the relationship associated with stratifying patients according to their predicted risk of death or dependency in the absence of treatment: those at the highest predicted risk from either event stand to experience the largest absolute benefit from alteplase with no indication of harm amongst those at lower predicted risk. It is concluded that prediction models for harmful side-effects based on simple clinical variables measured at baseline in randomised trials appear to offer little use in targeting treatments. Better separation between harmful events like bleeding and overall poor outcomes is required. This may be possible through the identification of novel (bio)markers unique to haemorrhage post treatment.

List of Publications and Presentations

Papers in peer reviewed journals

1. Thompson DD., Murray GD., Candelise L., Chen Z., Sandercock PAG., & Whiteley WN. "Targeting aspirin in acute disabling ischemic stroke: an individual patient data meta-analysis of three large randomised trials" International Journal of Stoke, 2015, DOI: 10.1111/ijs.12487

2. Thompson DD., Lingsma HF., Whiteley WN., Murray GD., & Steyerberg EW. 2014 "Covariate adjustment had similar benefits in small and large randomised controlled trials". Journal of Clinical Epidemiology, 2014, doi:10.1016/j.jclinepi.2014.11.001

3. Whiteley, W.N., Thompson, D., Murray, G., Cohen, G., Lindley, R. I., Wardlaw, J. & Sandercock, P. "The effect of alteplase within 6 hours of acute ischemic stroke on all-cause mortality (3rd International Stroke Trial)", Stroke, 2014; 45: 3612-3617

4. Thompson DD., Murray GD., Sudlow CLM., Dennis M., & Whiteley, W. 2014, "Comparison of statistical and clinical predictions of functional outcome after ischemic stroke", PLoS ONE 9(10): e110189. doi:10.1371/journal.pone.0110189

5. Thompson DD., Murray GD., Sudlow CLM., Dennis M., & Whiteley, W. 2014, "Formal and informal prediction of recurrent stroke and myocardial infarction after stroke: a systematic review and evaluation of clinical prediction models in a new cohort", BMC Medicine, 12:58

6. Whiteley, W.N., Thompson, D., Murray, G., Cohen, G., Lindley, R. I., Wardlaw, J. & Sandercock, P. 2014. "Targeting Recombinant Tissue-Type Plasminogen Activator in Acute Ischemic Stroke Based on Risk of Intracranial Hemorrhage or Poor Functional Outcome: An Analysis of the Third International Stroke Trial", Stroke, 45, 1000-1006

Conference proceedings

1. ISCB35 2014: (Oral presentation) “Covariate adjustment has similar benefits in small and large randomised controlled trials”. Thompson DD., Lingsma HF., & Steyerberg EW.

2. CTMC 2013: (Oral presentation) “The risks and benefits of RTPA in acute ischemic stroke for patients at high risk of intracranial haemorrhage and poor functional outcome: a secondary analysis of the IST-3 trial and systematic review of prediction models”. Thompson D., Murray G., & Whiteley W. Trials 2013, 14(Suppl 1):O9 (29 November 2013)

3. CTMC 2013: (Oral presentation) “A novel measure of treatment benefit for an ordinal scale: a case study of the IST-1 and the IST-3 stroke trials”. Thompson D., Whiteley W., & Murray G. Trials 2013, 14(Suppl 1):O48 (29 November 2013)

4. CTMC 2013: (Oral presentation) “Prediction of recurrent stroke and myocardial infarction after stroke: a systematic review of clinical prediction models”. Thompson D., Murray G., & Whiteley W. Trials 2013, 14(Suppl 1):O76 (29 November 2013)

5. CTMC 2013: (Poster presentation) “Predicting risks and benefits of treatment with aspirin in the acute stage of ischaemic stroke: an analysis of 3 large randomised controlled trials”. Thompson D., Murray G., & Whiteley W. Trials 2013, 14(Suppl 1):P117 (29 November 2013)

6. ESC 2013: (Poster presentation) “The performance of prognostic models for predicting occlusive vascular events after stroke: a systematic review”. Thompson D.D., Murray G.D., & Whiteley W.N.

7. UKSF 2012: (Poster presentation) “The performance of prognostic models for predicting occlusive vascular events after stroke: a systematic review”. Thompson D.D., Murray G.D., & Whiteley W.N. International Journal of Stroke, vol 7, s2, p34

Tables

Figures

Abbreviations

AF	Atrial Fibrillation
AUROCC	Area Under the Receiver Operating Characteristic Curve
BI	Barthel Index
CAST	the Chinese Acute Stroke Trial
CI	Confidence Interval
CPH	Cox Proportional Hazards
CPM	Clinical Prediction Model
CT	Computerised Tomography
DBP	Diastolic Blood Pressure
DVT	Deep-Venous Thrombosis
ECG	Electrocardiogram
ESS	Edinburgh Stroke Study
FDA	Food and Drug Administration (US)
FGM	Fine and Gray Model
GCS	Glasgow Coma Scale
HR	Hazard Ratio
ICH	intracerebral haemorrhage
IPD-MA	Individual Patient Data Meta-Analysis
IQR	interquartile range

IS	Ischaemic Stroke
IST-1	the first International Stroke Trial
IST-3	the third International Stroke Trial
KM	Kaplan-Meier
MAST-I	the Italian Multicentre Acute Stroke Trial
MI	Myocardial infarction
MRI	Magnetic Resonance Imaging
mRS	modified Rankin Scale
NIHSS	National Institutes of Health Stroke Scale
NRI	Net Reclassification Index
OHS	Oxford Handicap Scale
OCSP	Oxfordshire Community Stroke Project
OR	Odds Ratio
ORC	ordinal *c*-index
PE	Pulmonary Embolism
POLR	Proportional Odds Logistic Regression
RCS	Restricted Cubic Splines
rtPA	recombinant tissue plasminogen activator
SAH	subarachnoid haemorrhage
SBP	Systolic Blood Pressure

SD	Standard Deviation
SE	Standard Error
SRS	Simple Random Sample
SSV	Six Simple Variables
sHR	subdistribution Hazard Ratio
TIA	Transient Ischaemic Attack
VISTA	Virtual International Stroke Trial Archive

Chapter 1: Background and Introduction

Background and summary

How are the results from individual trials used to make treatment decisions for individual patients? How applicable is the average treatment effect to those at a very low or a very high risk of 'bad outcome'? Such questions are of utmost importance to both the clinician and the patient – especially when harmful side-effects are associated with treatment. This chapter outlines the argument that the targeted treatment of patients based on the balance of predicted harms and benefits could yield overall improvements in patient recovery. The stratified treatment of stroke patients is investigated.

1.1 Introduction

The evidence-based medicine paradigm states that the best treatment decisions are made by a doctor on the behalf of his or her patients with the support of objective scientific evidence (Guyatt et al., 1992). This evidence will typically come from two or more randomised controlled trials each suitably designed to assess the efficacy of the experimental treatment (Senn, 2007). The body of available trial data for a given indication may be summarised through systematic review and meta-analysis providing a single source of information used to guide the application of any licensed treatments at the clinic level. However, the clinician is faced with a dilemma: how applicable is the evidence in the case of the individual patient? The clinician may wonder whether the individual patient is well represented by the *trial-patients* and might reflect on the inclusion and exclusion criteria as well as any presented demographic data to assess whether their patient is, in some sense, *interchangeable*. This is often referred to as the *external validity* of a trial. Indeed a treatment that works on average could result from treatment benefit amongst a small proportion of high risk patients (i.e., those with a high control event rate) and no treatment effect,

or even a harmful treatment effect, amongst the vast majority of low risk patients (i.e., those with a low risk control event rate) (Kent and Hayward, 2007a). A primary trial publication presents the main findings but will frequently include a plethora of subgroup analyses thus acknowledging the underlying heterogeneity in the patient population. Pocock *et al*. summarised the various statistical concerns associated with this approach citing inadequate statistical power and uncontrolled type I error rates (Pocock et al., 2002).

If there is the suggestion that a given treatment works for some patients but not for others then how should investigation proceed? The pooling of aggregate results from subgroup analysis across trials is possible, but is dependent upon the availability of the relevant data (Thompson and Higgins, 2005). The sharing of the original patient level data enables better use of existing data resources (e.g., defining events, handling missing data, adjusting for variables etc. all in a unified manner). This has the potential to yield more reliable results thanks to the level of consistency that can be achieved (Riley et al., 2010). Since 2000 there has been a substantial rise in the number of Individual Patient Data Meta-Analyses (IPD-MAs) with Riley *et al*. noting an average of around 49 publications per year from 2005 onwards – this suggests some positive movement since Andrew Vickers commentary on his personal struggles in obtaining IPD (Riley et al., 2010, Vickers, 2006). It is rarely the case that individual sub-group analyses will achieve a sufficient degree of discrimination which can uniquely characterise patients with regards to either treatment benefit or harm. Multivariable regression techniques which combine multiple risk factors predictive of outcome have greater statistical power, beyond that of the standard subgroup tests (Hayward et al., 2006). This is a far superior and more efficient way of exploring whether treatment effect varies with patient heterogeneity (either in prognosis or risk from harmful side-effects) (Kent et al., 2010, van der Leeuw et al., 2014, Pocock et al., 2014)

With access to individual patient data the opportunity to explore specific secondary questions relating to harmful treatment side-effects with greater statistical power is therefore possible. An article by Whiteley *et al*. is an excellent example of this approach (Whiteley et al., 2013). The authors present an investigation of the

clinically plausible interaction between the predicted risk of haemorrhage (harm), the predicted risk of thrombosis (benefit) and treatment with heparin across five individual trial datasets. Stroke is a heterogeneous condition with very few available drug treatments. Greater benefit may be possible by better targeting those existing treatments to those most likely to benefit. This thesis aims to explore this question in detail reflecting on available trial datasets.

1.2 Stroke epidemiology

Stroke is the disruption of blood flow to the brain caused by a either a blockage (ischaemic stroke) or a rupturing of a blood vessel (haemorrhagic stroke). This disruption starves the brain of oxygen and is typically expressed by visible functional disability, e.g., limpness, slurred speech etc. Stroke is a leading cause of disability. It is currently ranked as the fourth most common cause of death and, despite a reported decrease in its incidence across developed countries, the absolute number of strokes continues to rise as a result of an ever aging population (Go et al., 2014). The most commonly used definition of stroke was provided by the World Health Organization (WHO) in 1976 (Hatano, 1976). They defined a stroke as the:

> "…rapidly developing clinical signs of focal (at times global) disturbance of cerebral function, lasting more than 24 h or leading to death with no apparent cause other than that of vascular origin."

The mechanisms through which cerebral blood flow is disrupted can be uniquely characterised as being either: ischaemic or haemorrhagic in origin. The advent of brain imaging has enabled the reliable distinguishing of the two (Donnan et al., 2008). The majority of strokes that occur are ischaemic and make up 80-85% of all strokes (Rathore et al., 2002). Ischaemic strokes are caused by an obstruction in a blood vessel which results in the under-perfusion of brain tissue. This obstruction can arise via: (i) embolism (either arterial or cardiac in origin); (ii) decreased perfusion (from stenosis – a narrowing of the arteries); or (iii) thrombosis (blood clotting formed in the heart or the blood vessels) (Mohr et al., 1997). Haemorrhagic strokes make up the remaining 15-20% of strokes and can be further categorised as being either subarachnoid (SAH) or intracerebral haemorrhages (ICH).

Historically suspected stroke cases which resolve within 24 hours from symptom onset are termed Transient Ischaemic Attacks (TIA) (Albers et al., 2002). More recently this categorisation of the duration of symptoms has come into question and a tissue-based definition has been advocated (Sacco et al., 2013). This incorporates the presence or absence of evidence of infarction via brain imaging, indicating that strokes can indeed occur within a 24 hour period. Brain ischaemia may be best viewed then as a continuum with both duration and the extent of infarction having serious implications for patient recovery (Easton et al., 2009).

Stroke places a considerable burden on health care. A study comparing the total health care expenditure on stroke at an international level suggested that the cost of stroke across eight countries constituted approximately 3% of spending (Evers et al., 2004). The societal cost of stroke in the UK has been estimated as being around £8.9 billion per year whilst in the US the cost has been estimated as $20.6 billion per year (Saka et al., 2009, Go et al., 2014).

1.2.1 Stroke outcomes

1.2.1.1 Functional outcome following a stroke

There are a number of ways to measure functional ability following a stroke. Each approach places a patient on some ordinal scale of severity or dependency providing a qualitative assessment of disability after stroke. These scales are multi-faceted incorporating a spectrum of possible stroke related deficits e.g.: mobility; consciousness; and responsiveness. A single score for a given patient is determined via patient interview by a trained professional (e.g., a stroke nurse or physician) at an appropriate time point.

The National Institutes of Health Stroke Scale (NIHSS) summarises 15 unique elements of impairment scoring patients on a scale that ranges from 0 to 42 in unit increments of worsening severity (Lyden et al., 1999). When measured at baseline it indicates the severity of the initial stroke which is useful for establishing the likely prognosis. When recorded later in follow-up it can be used as a primary outcome in trials. The Scandinavian Stroke Scale (SSS) summarises nine unique elements of impairment and ranges from 0 to 58, it acts in the opposite direction to the NIHSS

meaning that lower SSS scores are associated with more severe strokes (Scandinavian Stroke Study Group, 1985). As is the case with the NIHSS, the SSS can be utilised as a measure of severity for the initial stroke and for describing longer-term functional outcome. The NIHSS and the SSS largely measure the same attributes and have a particularly strong correlation. A conversion model has been previously described enabling the translation of the NIHSS to the SSS and vice versa (Gray et al., 2009). The Barthel Index (BI) measures ten activities of daily living covering aspects regarding ability to carry out personal tasks (for example, toilet use, grooming etc.) as well as mobility (Mahoney and Barthel, 1965). The BI measure ranges from 0 to 100 and is used as a measure of functional outcome following stroke – lower scores indicate more dependency. The modified Rankin Scale (mRS) is a seven level ordinal score of disability ranging from 0 (fully recovered) to 6 (dead) in unit increments of increasing dependency (van Swieten et al., 1988). A review by Quinn *et al*. found that the mRS was the most commonly used primary endpoint in randomised stroke trials (Quinn et al., 2009). Another scale, with the same number of levels as the mRS where each of the corresponding levels has a qualitatively similar definition to the mRS, is the Oxford Handicap Scale (OHS) (Bamford et al., 1989). The OHS was developed for application in a community setting with an emphasis on reporting patient handicap resulting from a stroke (New and Buchbinder, 2006). For the most part the mRS and the OHS are interchangeable.

Goldie *et al*. identified strong correlations amongst the most commonly used stroke outcome scales (i.e., NIHSS, mRS, SSS, and BI) to an extent which was much higher than was previously reported (Goldie et al., 2014)

1.2.1.2 Associated complications in recovery

After a stroke, patients are at greater risk from various intermediate events and complications (Langhorne et al., 2000). Pneumonia due to aspiration is common amongst stroke patients (Armstrong and Mosher, 2011). In one study conducted at a hospital in Glasgow, Scotland the rate of pneumonia after stroke was around 20% (Sellars et al., 2007). Post-stroke depression also occurs at a high rate, with about a third of all stroke sufferers experiencing depression (Hackett et al., 2005). Stroke patients aged 65 and above are at a heightened risk from stroke recurrence with

around 25% suffering a recurrence by five years (Mohan et al., 2011, Go et al., 2014). A systematic review of 39 separate studies suggests that the risk of myocardial infarction can range from 0.5% to 5.7% (Touzé et al., 2005). A Canadian multicentre stroke registry suggested a 1% risk of pulmonary embolism after stroke (Pongmoragot et al., 2013).

This list is not exhaustive but serves to illustrate the volume of 'bad-outcomes' that can result from a stroke. Early complications in the aftermath of a stroke are associated with additional increased risk of mortality and poor functional recovery beyond that of those risk factors recorded at baseline (Grube et al., 2013).

1.2.2 Risk factors

There are a number of known prognostic risk factors for the primary occurrence of stroke as well as recurrent events (Sacco et al., 1997, Warlow et al., 2003, Go et al., 2014). A prognostic risk factor is a measure which is associated with the occurrence of some clinical outcome or endpoint. Risk factors can be *modifiable* or *fixed*. Both are important in understanding patient prognosis but those modifiable risk factors are generally regarded as being of particular interest for targeted intervention (Riley et al., 2013).

1.2.2.1 First ever stroke

Fixed risk factors for ischaemic stroke are unchangeable either because they are intrinsic to the patient (e.g., gender or race) or inevitable (e.g., increasing age). Potentially modifiable risk factors include: hypertension (high blood pressure); atrial fibrillation; TIAs; myocardial infarction; carotid artery disease; cigarette smoking; diabetes mellitus; and obesity (Go et al., 2014). Controlling these risk factors through public education (e.g., helping smokers quit) or treatment intervention (e.g., treating hypertension) is of upmost importance. This can yield considerable health gains for the individual patient as well as offering economic gains for an often overly stretched health care budget (Donnan et al., 2008, Bornstein et al., 2006, Shah and Cole, 2010). Risk factors for SAH include: a history of hypertension, smoking and excessive alcohol consumption (Ruigrok et al., 2001). Similarly, for ICH risk factors

include: increasing age, hypertension, male gender and excessive alcohol consumption (Ariesen et al., 2003).

1.2.2.2 Recurrent stroke

Stroke recurrence is common amongst first-time stroke patients but rates recorded within the control arms of stroke prevention trials have reduced over time from 9% in 1970 to 5% in 2000 (Go et al., 2014). Mohan *et al*. used multivariable regression techniques to identify risk factors associated with recurrence over a 10 year follow-up period within a South London Stroke Register. They found that the presence of atrial fibrillation, myocardial infarction, and hypertension were all associated with recurrent stroke (Mohan et al., 2009). In a similar study Hillen *et al*. found that both atrial fibrillation and diabetes mellitus were associated with recurrence in follow up (Hillen et al., 2003).

In general, overall patient recovery following a first time stroke depends upon various characteristics associated with aging (i.e., additional co-morbidities, for example, heart diseases and dementia) as well as the impact of the initial stroke (i.e., high NIHSS scores) (Appelros et al., 2003).

1.2.3 Treating acute ischaemic stroke patients

The aim of treatment as early as possible after the onset of an acute ischaemic stroke is to restore normal blood flow to the affected area of the brain and therefore minimise the amount of damage caused. There are two effective medications: aspirin; and recombinant tissue plasminogen activator (rtPA).

1.2.3.1 Aspirin

Aspirin is an antiplatelet drug which, if given within 48 hours from acute ischaemic stroke onset, reduces the risk of death or dependency by 1% (Sandercock et al., 2008). Platelets are responsible for the formation of blood clots: antiplatelet treatments inhibit this process and counteract the formation of new clots. Although aspirin is associated with only a small absolute reduction in death or dependency, its effect is still important. It is cheap, easy to administer and applicable to a large proportion of stroke patients (Gilligan et al., 2005). Risk of bleeding may be of

concern but rarely occurs with an increased risk of about 2 for every 1000 patients treated (Chen et al., 2000). Risk of bleeding is still an important consideration though and is one of the main contraindications to starting treatment (e.g., a previous gastrointestinal bleed or high blood pressure).

1.2.3.2 Recombinant tissue plasminogen activator

Thrombolysis with rtPA is effective in the treatment of vascular conditions like myocardial infarction and in recent years has emerged as an effective treatment for acute ischaemic stroke (The GUSTO investigators, 1993, White and Van de Werf, 1998, Wardlaw et al., 2012a). Treatment with rtPA within 3 hours of acute ischaemic stroke onset increased patients chance of a good recovery by six months (mRS≤2) by approximately 4% (Wardlaw et al., 2012a). Instead of preventing further clotting, rtPA works by dissolving the original clot thus restoring normal blood flow (Yaghi et al., 2014). Time to treatment with rtPA after acute ischaemic stroke onset is of vital importance as there is empirical evidence to suggest that the benefit of rtPA decreases over a four and a half hour window from onset (Lees et al., 2010). Patients treated with rtPA are at risk from suffering symptomatic intracranial haemorrhage (SICH) which frequently lead to premature death (Wardlaw et al., 2012a). Various definitions have been adopted in the identification of SICH post rtPA through randomised trials and observational registries. An SICH as defined according to the National Institute of Neurological Diseases and Stroke (NINDS) trial attributes *any* deterioration of the patient's condition as measured by the NIHSS within 36 hours of thrombolysis plus the development of an intracerebral haemorrhage (ICH) verified by brain imaging (i.e., CT or MRI) (The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group, 1995). The European Cooperative Acute Stroke Study (ECASS) II defined an SICH as any type of visible bleed during post-treatment brain imaging as associated with clinical deterioration or a neurological deterioration of four or more NIHSS points within seven days of starting thrombolysis (Hacke et al., 1998). Finally, the SITS-Monitoring Study (SITS-MOST) defined SICH as a local or remote type 2 parenchymal haemorrhage identified on CT taken between 22 and 36 hours from the administration of thrombolysis in addition to neurological deterioration of four or more NIHSS points

baseline or the lowest score after baseline recorded up until seven days or leading to death (Wahlgren et al., 2007).

1.3 Stratified Medicine

Stratified medicine is the targeted treatment of groups of individuals who test *positive* for some measureable trait (e.g., a biomarker or a combination of multiple characteristics via statistical modelling) who experience a differential treatment effect. Hingorani *et al*. distinguish aspects pertaining to stratified treatment decisions on an absolute risk scale (i.e., assuming a constant relative treatment effect) or on a relative risk scale where the relative effect of treatment necessarily differs (Hingorani et al., 2013).

In the absolute risk case, if the overall relative treatment effect is constant over patient prognosis then those with the poorest prognosis stand to gain the largest absolute risk reduction in contrast to those with a good prognosis (Rothwell, 2007a). In statistical terms this corresponds to the assumption that that there is no interaction between predicted prognosis and the effect of treatment on outcome. If this holds true then it makes sense to interpret the gain in treatment benefit with respect to the risk of treatment harm as illustrated in Figure 1-1 (Glasziou and Irwig, 1995). This would then suggest that a 'breakpoint' for net benefit conditional on baseline risk exists. It is important to understand the risk of harm and establish the threshold at which benefit from treatment exceeds harm. Note that a constant rate of harm is assumed in Figure 1-1 though there is no reason why this could not also depend on the baseline risk and also have a positive slope (van der Leeuw et al., 2014).

A non-constant relative treatment effect invalidates the simple risk-benefit model. This would then correspond to a stratified approach with the relative treatment effect differing according to one or more identifiable subgroups of a patient population. Any assessments of interaction should though be based upon biological grounds in advance to testing. Validation and replication of biological interactions are a central aspect of stratified medicine (Hingorani et al., 2013).

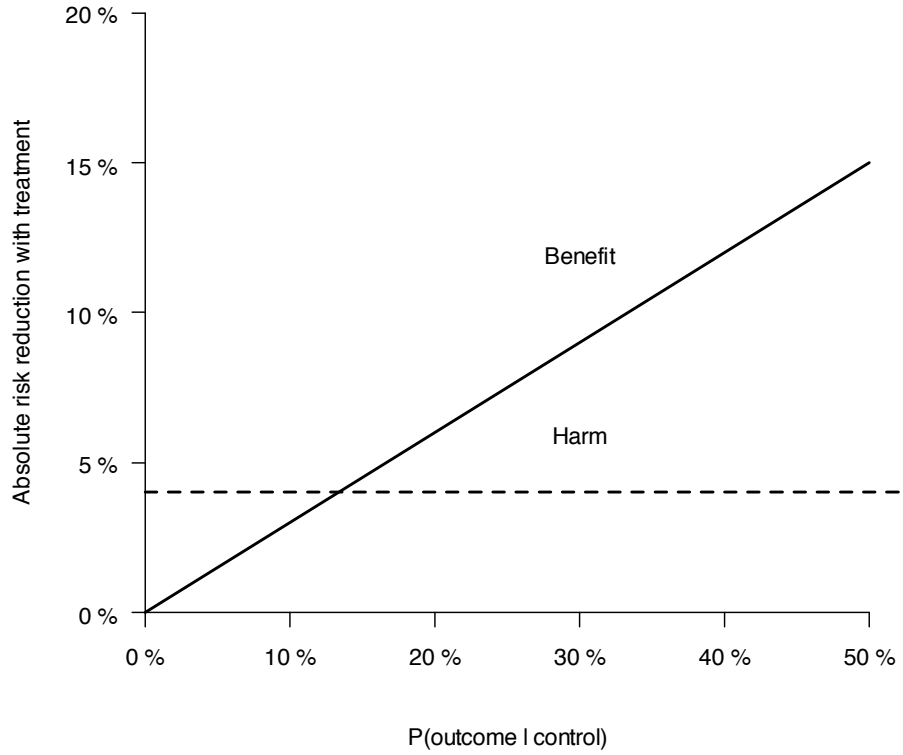


Figure 1-1 Basic risk-benefit decision model concept

1.3.1 Biomarkers

The WHO define a biomarker as (Strimbu and Tavel, 2010):

> "…any substance, structure, or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease."

In this sense any patient measure associated with an outcome of interest is a 'biomarker', though more often than not so called *laboratory-measured* biomarkers are of primary research interest. An example of such a biomarker is the predictive Human Epidermal growth factor Receptor 2(HER-2) marker for breast cancer patients (Hudis, 2007). Studies have shown that those who test positive for the HER-2 marker may be treated with trastuzumab (a monoclonal antibody) whilst those that test negative may not. A recent systematic review of the use of biomarkers within drug licenses issued by the European Medicines Agency (EMA) suggested that the frequency of their use in practice did not entirely match with the research efforts and expenditure that biomarker research receives – translation from the lab bench to the

prescription slip has not been as productive as hoped (Malottki et al., 2014). However, these licenses were predominantly for HIV and cancer treatments and focused more on efficacy than toxicity. Few biomarkers though make it into clinical practice. Of those published biomarkers in cancer under 1% have any clinical utility (Kern, 2012).

1.3.2 Prediction models

When multiple risk factors are associated with patient prognosis, e.g., recurrent stroke or poor functional outcome after acute stroke, it is sensible to consider the impact of all clinically important variables at once thus creating a detailed picture of an individual's risk profile (Kent et al., 2010). A clinical prediction model formally combines multiple variables for the prediction of some clinical endpoint in a given population of patients (Steyerberg, 2009, Steyerberg et al., 2013). Each variable is weighted by an estimated coefficient which reflects the importance of a unit increase or decrease of that variable upon the risk/chance of outcome. When multiple risk factors are predictive of treatment response then a risk-stratified approach may be possible. An excellent illustration of this is found in the study by Farooq *et al*. who demonstrated that treatment decisions for patients with complex coronary artery disease can be guided by unique prediction models (Farooq et al., 2013). The authors extended an existing prediction model (the SYNTAX score), which original comprised of just anatomical features, by including clinical variables. The SYNTAX II model provides predicted risks for patients four year mortality if treated with percutaneous coronary intervention (PCI) and separately the risk when treated with coronary artery bypass graft surgery (CABG). Patients and doctors may then make judicious decisions as to which treatment yields the largest potential benefit.

The statistical methodology required is neither new nor challenging – indeed sensible statistical tests of the additivity assumption are well established as are the proper approaches to model development (Harrell, 2001). Despite this, efficient quantitative methods are rarely utilised in practice even though they have the potential to improve the translation of trial findings to individual patients (Pocock et al., 2014).

The methodological aspects central to the development and assessment of clinical prediction models will be discussed in Chapter 2.

1.3.3 The targeted treatment of stroke patients

Acute ischaemic stroke is a heterogeneous condition with very few effective treatments. Those few available treatments come with benefits and harms. The *one-at-a-time subgroup* approach is not an efficient way of exploring the underlying heterogeneity of risk of harm or chance of benefit. A multivariable risk prediction approach has the potential to yield substantial gains in overall clinical benefit (Kent et al., 2010). For instance, not everyone is eligible for rtPA though a risk stratified approach could help elucidate whether clinicians are under– or over-treating their patients based on perceived risk factors from SICH. Additionally, a risk stratified approach could help yield greater benefits from treatment with aspirin where rtPA is not an option.

An approach based on risk prediction is reliant upon models achieving a sufficient level of discrimination so as to ensure the accurate targeting of treatments. Using trial datasets this approach will be explored in this thesis is to explore the possible risk stratified treatment of stroke patients weighting by the predicted risk of harm and the chance of outcome in the absence of treatment.

1.4 Thesis outline

The methods used in the development and evaluation of a clinical prediction model are introduced and discussed in Chapter 2. This chapter covers various aspects of methodology which recur throughout the course of this thesis and may therefore be used by the reader for reference. A systematic review of clinical prediction models for the prediction of recurrent ischaemic stroke or myocardial infarction following an ischaemic stroke is presented in Chapter 3. Those identified models are then evaluated in a new prospective single centre hospital cohort of stroke patients in Chapter 4. Their relative performance is compared to that of the treating clinicians' who each made informal predictions of vascular endpoints for their patients. In Chapter 5 a similar question is explored, evaluating a number of existing prediction models for the prediction of poor functional outcome. These models are chosen from a previously conducted systematic review and again tested within the same dataset used in Chapter 4 (Veerbeek et al., 2011). These data also contain clinicians' informal predictions of functional outcome allowing a further comparison of informal versus formal methods of prediction.

Various trial datasets are re-analysed as part of this thesis with the results presented in Chapters 6 through to 9. Chapter 6 is an IPD-MA of three acute ischaemic stroke trials of aspirin. Here a balance between predicted risk of treatment harm and treatment benefit favouring good functional outcome is sought. Chapter 7 presents an analysis of a recent rtPA trial to investigate whether the effect of rtPA on the reduction of six month death or dependency varies with predicted risk of treatment harm. Using the same dataset Chapter 8 investigates the possible long-term impact of rtPA on mortality and explores whether this depends upon patients prognosis or delay to treatment. Finally, in Chapter 9 data from a large stroke trial archive are used to investigate whether the separate impact from early thrombotic and early haemorrhagic events have different associations with the risk of mortality.

Chapter 2: The development and evaluation of a clinical prediction model

Background and summary

Methods for the development of clinical prediction models and the metrics used to assess their performance are discussed. These concepts are illustrated using worked examples based on stroke trial datasets.

2.1 Introduction

Medical data collected through routine observation or direct experimentation is opaque and uninterpretable without statistical examination. Underlying patterns and relationships can be explored by first positing some plausible association between an outcome and one or more additional measurements. The *existence* of an underlying rule or model involves the combination of a number of associations each describing the conditional relationship between outcome and predictors. The aim of this chapter is to discuss and illustrate various methodological aspects of model development and model evaluation. The methods introduced here will be used throughout this thesis.

2.2 What is a clinical prediction model?

A statistical model is a description of some phenomenon of interest investigated through experimentation and numerical observation (Dawid and Senn, 2011). Cox highlighted two broad types of statistical model: first the *substantive model*, the form of which stems from some scientific or biological justification; and second the *empirical model*, a 'black-box' approach commonly adopted in regression modelling where data are used to estimate effects (Cox, 1990). There are two general motivations for modelling data. Firstly, they can provide an *explanation* of the underlying mechanisms perhaps with some adjustment for potential confounders. Secondly, assuming that the model is generalisable, they can be used to make *predictions* for future subjects. This framework is applicable to any experiment

where quantitative data have been collected. For example, consider the Framingham Risk Score (FRS) derived from a population of men and women free of Coronary Heart Disease (CHD) who were prospectively followed up for 12 years for the prediction of 10 year CHD risk. The FRS is based on six variables: age, total cholesterol, smoking status, high-density lipoprotein cholesterol and systolic blood pressure (Wilson et al., 1998). The most up-to-date version of the FRS has been adapted to predict Cardiovascular Disease (CVD). The FRS plays an important role in the prescription of statins used in the primary prevention of CVD, where statins are recommended to those with a predicted risk of 20% or greater of developing CVD (National Institute for Health and Care Excellence (NICE), 2006).

A Clinical Prediction Model (CPM) is therefore a combination of appropriately weighted clinical risk factors associated with a clinical outcome of interest. The proposed fit may be used to either explain the underlying associations or else to predict patient risk. The focus throughout this thesis is predominantly on the latter function.

2.3 Simplicity versus complexity

One way of quantifying the complexity of a proposed statistical model is by considering the number of included parameters. Statisticians denote this as the 'degrees of freedom' which are the number of required parameter estimates for each outcome-predictor association, i.e., for continuous variables and dummy variables (as only *contrasts* are identifiable). The degrees of freedom summarise all included effects, i.e., main effects, interactions and any non-linear transformations. Strictly speaking, formal and informal tests used to screen variables for inclusion or exclusion should also incur additional expenditure of the degrees of freedom, however, this is rarely acknowledged in practice. A useful analogy is to view this as a financial transaction: the degrees of freedom act as currency; the data denote the available balance; and the goodness-of-fit is the service provided. With access to all of this information the wisdom of the investment can be assessed. In general, the model which strikes the most economical balance between the degrees of freedom spent and the goodness-of-fit to the data is the best value. In order to avoid

bankruptcy it is necessary to consider whether the data permit the proposed model. In practice the data will frequently impose some (often major) limitations on the development of a model.

Model complexity should be tuned to meet the intended study purpose. It has been suggested that the level of detail that a complex model provides is better suited when *explanation* is the goal, for example in effect estimation or hypothesis testing (Núñez et al., 2011). Investigators can include multiple confounders in an attempt to extract some of the noise and heterogeneity from the system. In the case of continuous outcome data this produces a more precise estimate, whilst in the case of a binary outcome a more subject specific effect estimate is obtained (Robinson and Jewell, 1991). The resulting effect estimate may be more informative and the hypothesis test more reliable. Model overfitting is an important consideration and the sample data details the degree of plausible complexity. In the case of *prediction*, generalisability of the model fit is the primary motivation (Steyerberg, 2009). The predicted risks estimated on the development population must be generalisable to a different population to serve some purpose. This generalisability becomes more plausible when the proposed fit conditions less on the observed data and thus avoids capitalising on certain *data-specific-quirks*. This is one reason why a simple or parsimonious model is often preferred. Simple models which require little manipulation from the user, with only a few easy to measure input variables, will have superior face validity making it more likely that the proposed model is adopted (Moons et al., 2012b).

Throughout this thesis simplicity will frequently be preferred over complexity.

2.4 Model development

The focus of this chapter is on the methods used in the development and evaluation of a clinical prediction model. The methods introduced here are used throughout this thesis. Study design will be touched on only briefly with greater detail deferred to Chapter 3.

2.4.1 The general framework

The development of a clinical prediction model should follow a sensible procedure or protocol defined from the outset (Collins, 2011). First, a relevant question of clinical importance must be asked, for example: *what is the six month risk of death or dependency following an acute ischaemic stroke*? To answer such a question, reliable high quality data must be used, e.g., patients recruited consecutively, who are prospectively followed-up recording all events of *a priori* interest over a fixed period of time. A Randomised Clinical Trial (RCT) of acute ischaemic stroke patients comparing some active treatment to a control will fit most of these desirable criteria, although may suffer from poor generalisability especially if the trial followed strict exclusion and inclusion criteria. If existing data sources are not available or inadequate then new data should be collected for the purpose of model development. When selecting which risk factors to include it is important to consider whether there is a body of existing knowledge available, this should be systematically reviewed so that clinically important risk factors – regardless of their statistical significance – can be included (Moons et al., 2012b).

It then follows that the performance of the proposed model is scrutinised within some external dataset, for example within a similar randomised trial run in a different country (Moons et al., 2012a). Issues relating to the selection of risk factors and evaluation will be discussed later in this chapter.

2.4.2 The logistic regression model

In this section some common approaches to analysing categorical data are discussed: in particular binary and ordinal outcome data.

2.4.2.1 Binary outcomes

Often statistical analysis will be carried out on outcome data, Y, which are strictly binary in nature. Frequently the end point of interest will be the presence or absence of some event, for example, the patient is alive or dead. Associations between the outcome and other measured patient characteristics can be estimated using logistic regression. Logistic regression assumes that a binomial, $Bin\ (1, \theta)$, outcome random

variable, Y, is dependent upon a set of observed covariates (or predictors), X_j, ($j = 1, 2, \ldots, p$). The probability of response for the i^{th} individual is then modelled as:

$$P(Y_i = 1 | \alpha, \boldsymbol{\beta}; \boldsymbol{X}_i) = \frac{\exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{X}_i)}{1 + \exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{X}_i)} . \quad \textbf{(2.1)}$$

The probability of a response is given by $\pi_i = \mathrm{P}(Y_i = 1 | \alpha, \boldsymbol{\beta}; \boldsymbol{X}_i)$ and similarly the probability of a non-response by $1 - \pi_i$. The probability for the i^{th} observation depends directly on a vector of p independent covariates, $\boldsymbol{X}_i = [X_1, X_2, \ldots, X_p]$, with the unknown parameters $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_p]$ and intercept α.

The unknown parameters specified in equation (2.1) are estimated by maximising the likelihood function,

$$\prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1 - y_i} .$$

There is no explicit solution for the partial derivatives of this function, it is therefore necessary to use numerical approximation (e.g., Newton-Raphson) to compute each of the $p + 1$ possible derivatives of the *log*-likelihood.

The logistic function described by equation (2.1) restricts all predictions to the range $0 \leq \pi_i \leq 1$ giving a sigmoidal shape between $\boldsymbol{\pi}$ and $\boldsymbol{\beta}^T \boldsymbol{X}$, although a simple transformation enables a linear interpretation (see Figure 2-1). Under the *logit transformation* the risk factors are said to act linearly on the log odds scale (Dobson, 1991),

$$\ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \alpha + \boldsymbol{\beta}^T \boldsymbol{X}_i . \quad \textbf{(2.2)}$$

Here, a unit increase in the X_{ij} (where $j = 1, 2, \ldots p$ and $i = 1, 2, \ldots, N$) risk factor is interpreted as an increase ($+\beta_j$) or decrease ($-\beta_j$) in the log odds of response holding all other $p - 1$ covariates constant. By exponentiating the regression coefficient the

effect can be interpreted on the odds scale with the relevant odds ratio (OR) where the OR describes the increase (OR>1) or decrease (OR<1) in the odds of some outcome, $\pi / (1 - \pi)$, relative to some risk factor.

The logistic regression is the most common method for analysing binary data. One less commonly encountered method is the probit model, $\Phi^{-1}(\pi)$, where Φ is the cumulative probability function for the standard Normal. Historically the probit model has roots in dose response modelling (Dobson, 1991). This method will not be used in this thesis.

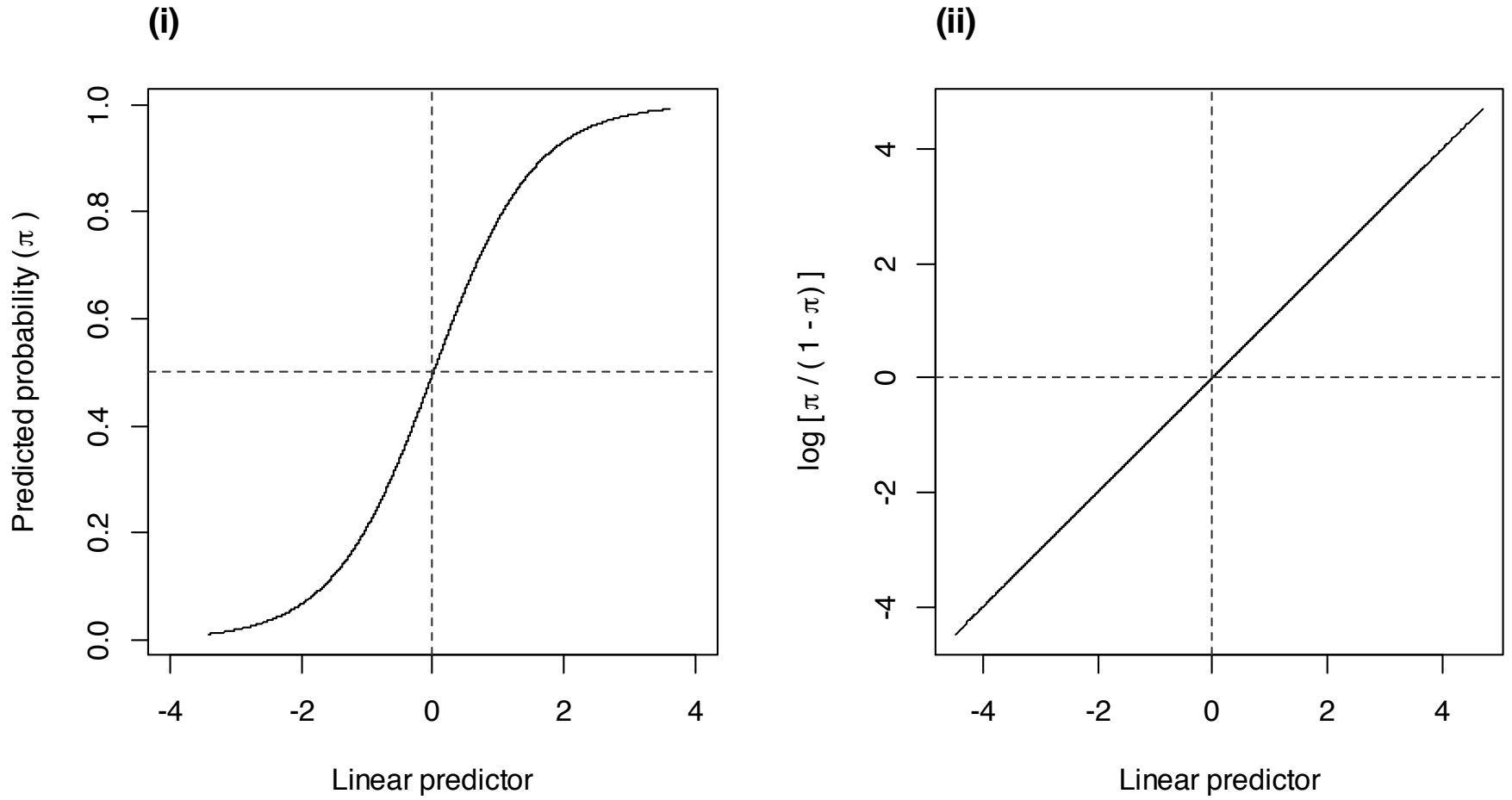


Figure 2-1 Example of the logistic function for a binary logistic regression showing the association between: (i) the linear predictor and the logistic function, where $0 \leq \pi_i \leq 1$; and (ii) the linear predictor and the logit transformation which has a linear interpretation on the log odds scale

2.4.2.2 Ordinal outcomes

Many medical conditions record patient disability at a fixed time point on an ordinal outcome scale. Such outcomes are commonly encountered in stroke and in head trauma studies. An ordinal outcome specifies a finite number of category levels (where $k > 2$) with the assumption that one level will qualitatively describe where a given patient belongs along the scale. In stroke these unique levels have a natural ordering specifying increasingly more severe levels of disability. It is common to see the modified Rankin Scale (mRS) used to score stroke patient disability. The mRS is a seven-point scale, ranging from fully recovered (a mRS of 0) to dead (a mRS of 6) in single unit increments. Investigators will often dichotomise the mRS enabling a binary interpretation, e.g., 0-2 versus 3-6, corresponding to 'alive and independent' and 'dead or dependent' respectively (Weisscher et al., 2008).

A more efficient use of the recorded information would be to analyse the entire range of the ordered response capitalising on the inherent ordinal structure. The proportional odds logistic regression (POLR) model achieves this by assuming that individual covariate effects are common across each of the $k - 1$ categorisations of the ordered outcome (McCullagh, 1980). Here the cut-point which splits the ordinal outcome into favourable versus unfavourable is moved across the ordinal scale creating all potential dichotomous splits. This invokes the proportional odds assumption which, if true, states that no matter how the outcome is dichotomised a constant (or near constant) effect will be obtained. This is a simple extension of equation (2.1):

$$P(Y_i \geq k \mid \alpha_1, \cdots, \alpha_k, \boldsymbol{\beta}; \boldsymbol{X}_i) = \frac{\exp(\alpha_k + \boldsymbol{\beta}^T \boldsymbol{X}_i)}{1 + \exp(\alpha_k + \boldsymbol{\beta}^T \boldsymbol{X}_i)}. \quad \textbf{(2.3)}$$

Here the regression parameters are assumed to be the same and constant for each of the k intercepts ($k = 1, \ldots, K$).

Again, a linear interpretation is obtained through the logit transformation with equation (2.2). By exponentiating the regression coefficients in the POLR model the so called common ORs are obtained. The relative risk of $Y_i \geq k$ is interpreted as

common across all potential cut-points of the ordered outcome. This is highlighted in Figure 2-2 which illustrates another interpretation of the proportional odds assumption as 'parallelism', that is for all response categorisations the corresponding predictors are parallel on the log odds scale.

The proportionality assumption can be assessed both formally and informally, although formal tests suffer from *extreme anticonservatism* (Harrell, 2001). Informal visual assessments of the residuals can be far more informative, e.g., examining plots of score residuals:

$$U_{ip} = X_{ip}(I(Y_i \geq k) - \hat{P}_{ik}) \quad \textbf{(2.4)}$$

Here $\hat{P}_{ik}$ is the predicted probability of the outcome $Y_i \geq k$ for the i^{th} individual, estimated using equation (2.3). $I(Y_i \geq k)$ is an indicator variable for the event predicted and X_{ip} is the i^{th} observation for the p^{th} predictor. The resulting 'forest-plot' of mean scores, $\overline{U}_{.p}$ with associated 95% CIs against the corresponding levels of Y can be inspected to determine how plausible the PO assumption is (Harrell et al., 1998).

This is illustrated in Figure 2-3 where any deviation away from the desirable flat relationship indicates a violation of the proportional odds assumption. However, others have shown that the model is fairly robust against such departures (McHugh et al., 2010).

Separate methods for the analysis of ordinal outcomes exist. The multinomial logistic regression model and the partial proportional odds logistic regression model are distinguished by the number of effect estimates obtained. This is reasonable when there is a clear violation of the proportional odds assumption such that multiple effects are associated with the possible categorisations of the ordinal outcome as a posed to just a single effect (Biesheuvel et al., 2008, Harrell, 2001). The drawback is a shift away from a simple interpretation toward a more complex and fragmented description; however, such a model may be necessary under certain conditions.

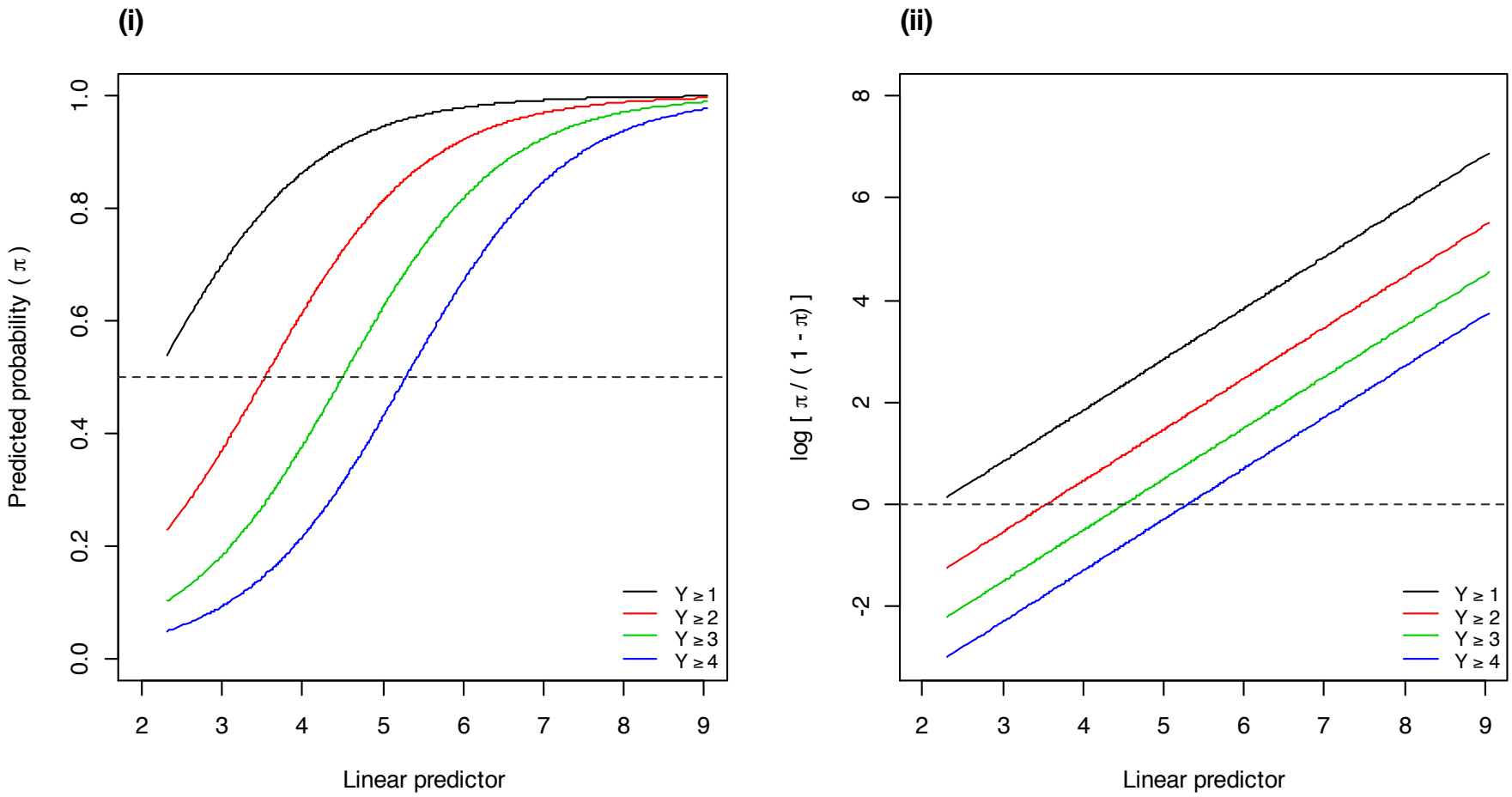


Figure 2-2 Example of the logistic function for a POLR showing the association between: (i) the $\beta^{T}X$ component of the linear predictor and the logistic function with $0 \leq \pi_{ik} \leq 1$; and (ii) the $\beta^{T}X$ and the logit transformation which has a linear interpretation

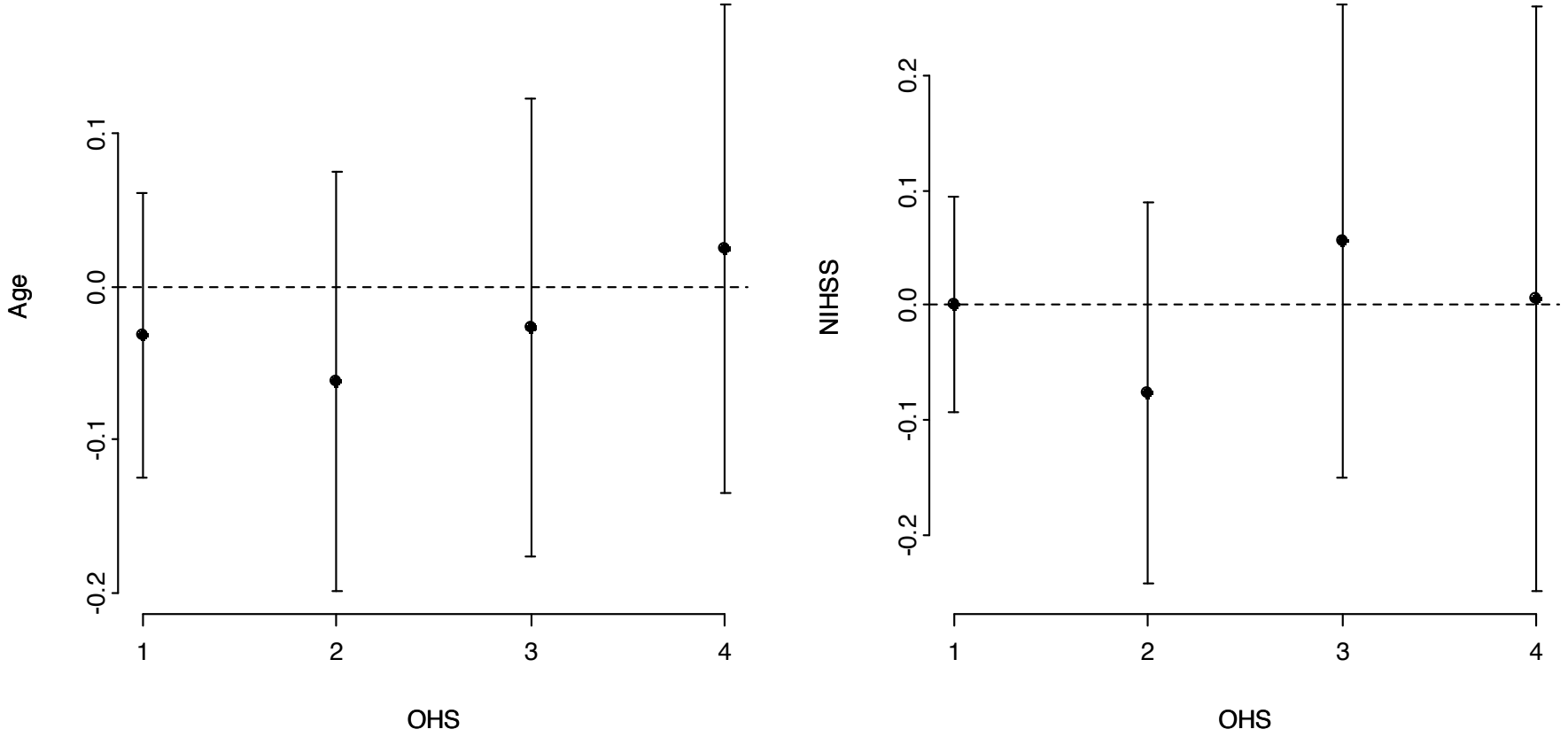


Figure 2-3 Example of a score residual plots from a multivariable POLR model fitting patient age and NIHSS

2.4.3 Selecting risk factors

Investigators are often faced with a multitude of potential predictors to include in their model(s). There are two schools of thoughts which may be used exclusively or perhaps in combination. The first argument is that investigators should only consider those predictors which appear in a pre-specified list of clinically important predictors, for example, those variables highlighted through a systematic review as *known* risk factors for poor functional outcome after stroke. The second approach is an automatic selection procedure whereby the *most important* relationships expressed by the data are chosen. Each approach places a premium on different forms of information: existing prior information; and newly obtained information (i.e., sample data). These are discussed below in reverse order.

2.4.3.1 Automatic variable selection

Stepwise variable selection follows an algorithm which iteratively selects variables until the best combination of predictors is reached. The goal is to produce an optimal fit which maximises some pre-defined stopping rule, e.g., the χ^2 test statistic with entry or removal at a P-value < α or else an improvement in the Akaikes Information Criterion (*AIC*). The algorithm can be set to choose predictors using: forward selection building from the intercept, i.e., the *null* model; backward selection, which starts with the full model and simplifies it by removing those 'unimportant' variables; or both forward and backward selection which involves both including and removing covariates in such a way that all combinations are considered. The main attraction to stepwise selection is that it is simple to implement and provides a quick and easy answer to the question: *which predictors should I include in my model*? However, provided that there is pre-existing information this is a pseudo-scientific approach. Several drawbacks make this method problematic (Steyerberg et al., 1999). Harrell summarised a number of these, citing: (i) biased small SEs; (ii) coefficients that are too large due to a selection process that favours the inclusion of effect sizes that are over-estimated; and (iii) the removal of intellectual input, which can often result in confusing models that exclude clinically important variables (Harrell, 2001).

2.4.3.2 Pre-specified variable selection

By using *a priori* clinical knowledge the problems that arise with data-dependent selection methods are avoided. One drawback is the requirement for the relevant background information to already exist. Under these circumstances automatic variable selection with a less stringent stopping rule (e.g., P-value < 0.10 or 0.20) may be used.

The generalisability of an estimated model can be assessed by considering the ratio of the number of events to the number of parameters to estimate. A simulation study suggested that a minimum of 10 events per parameter/variable (EPV) should be available as anything below this could result in overfitting and biased effect estimates (Peduzzi et al., 1996). For instance, consider a sample of 1000 acute ischaemic stroke patients with 10 years' worth of follow-up available by which point 100 of the patients suffer a recurrent stroke. Therefore in this example with 100 events, a minimum of 10 parameters can be reliably estimated. This is not a strict mathematical rule though and subsequent studies have suggested that this rule may even be relaxed to as low as 5 EPV (Vittinghoff and McCulloch, 2007).

In this thesis pre-specified risk factors and the 10EPV rule will be used throughout.

2.4.4 Model assumptions

Here three assumptions which apply to most regression models are considered (i.e., linear regression, Cox proportional hazards and logistic regression) (Harrell et al., 1988). First that the sample of patients studied is a random sample from the population of interest on which generalisable inferences can be made. Second, the functional relationship between predictor and response is a particular shape (e.g., linear, cubic, etc.). Finally, the predictors are additive, i.e., no interactions exist.

2.4.4.1 Random sample

Any statistical inferences drawn from a simple random sample of data from some well-defined population are unbiased and therefore generalisable to the population under study. The RCT is the 'gold-standard' for evaluating treatment efficacy which is then assumed to hold true of the *general patient population*. However, due to the

various exclusion and inclusion criteria this sample may bear little resemblance to the target population meaning that those recruited comprise of a highly selected subset of the target population (Senn, 2007). Despite this drawback datasets from RCTs offer some of the most robust and reliable data upon which to develop a model thanks to the prospective record of predictor variables as well as the ascertainment of adjudicated patient outcomes (Steyerberg, 2009). The validity of the prediction model is best assessed through model evaluation studies which are also informative about how generalisable the prediction model is. Design issues with model evaluation studies and the associated performance metrics will be discussed later on in this chapter.

2.4.4.2 Assessing linearity

The models presented in equations (2.1) and (2.3) implicitly assume linearity on the log odds scale. Irrespective of the scale, there is little reason why a continuous covariate, X, should have a linear relationship with $g(Y)$, other than when strong prior knowledge is available – where $g(\cdot)$ is some function of the outcome, e.g., the logit transformation. Non-linearity can be explored in a number of different ways. Cubic splines involve fitting cubic polynomial functions across designated intervals of a continuous covariate, X. These intervals are defined by the number and placements of *joins*, or *knots*, which span the range of X.

Consider placing three knots at A, B and C across X. With three cut-points four intervals are specified: $X < A$; $A < X \leq B$; $B < X \leq C$; and $C < X$.

$$f(X) = \beta_0 + \sum_{p=1}^{3} \beta_p X^p + \beta_4 (X - A)_+^3 + \beta_5 (X - B)_+^3 + \beta_6 (X - C)_+^3 \qquad \textbf{(2.5)}$$

Where $(u)_+ = u$ if $u > 0$ and equals 0 if $u \leq 0$. Note that accepting non-linearity here would imply rejecting the null hypothesis, H_0: $\beta_2 = \beta_3 = \ldots = \beta_6 = 0$.

In general, with k knots, there are $k + 3$ degrees of freedom used, excluding the intercept. By forcing the first and second derivatives of the function to be equal at the knots one can ensure a smooth continuous fit across the entire observed range of X.

One unfortunate property is instability in the tales of the distribution. The Restricted Cubic Spline (RCS) is advocated by Harrell, which, by forcing linearity in the tails, produce more stable predictions (Harrell, 2001). In addition, a saving is made in the number of parameters to estimate (reduced to $k - 1$, excluding the intercept),

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{k-1} X_{k-1} \tag{2.6}$$

where $X_1 = X$, and for $j = 1, \ldots, k - 2$,

$$X_{j+1} = (X - t_j)_+^3 - \frac{(X - t_{k-1})_+^3 (t_k - t_j)}{t_k - t_{k-1}} + \frac{(X - t_k)_+^3 (t_{k-1} - t_j)}{t_k - t_{k-1}}$$

The number of knots, not their location, is central to the quality of the fit, which is typically based upon percentiles of the distribution of the covariate (Stone, 1986).

In this thesis simpler transformations will be sought in place of the RCS fit provided that the improvement in fit is similar (e.g., a log transformation, or a quadratic expansion etc.). For example, in Figure 2-4 the quadratic fit is strikingly similar to the RCS fit with four knots, with a similar improvement (a model χ^2 of 12.85 compared to 12.82). The RCS can therefore be used as a powerful test for non-linearity. Although in this case a simpler fit is made by using a quadratic transformation which saves on a degree of freedom. The impact here is quantified by the size of the χ^2, this will be discussed in detail later in the chapter.

It is contradictory to advocate the use of hypothesis testing to select the functional form of a predictor whilst simultaneously oppose their use in variable selection. However, an elegant rebuttal to this is articulated by the following quote from Frank E. Harrell, Jr (Harrell, 2001):

> "Carefully fitting an improper model is better than badly fitting (and overfitting) a well-chosen one"

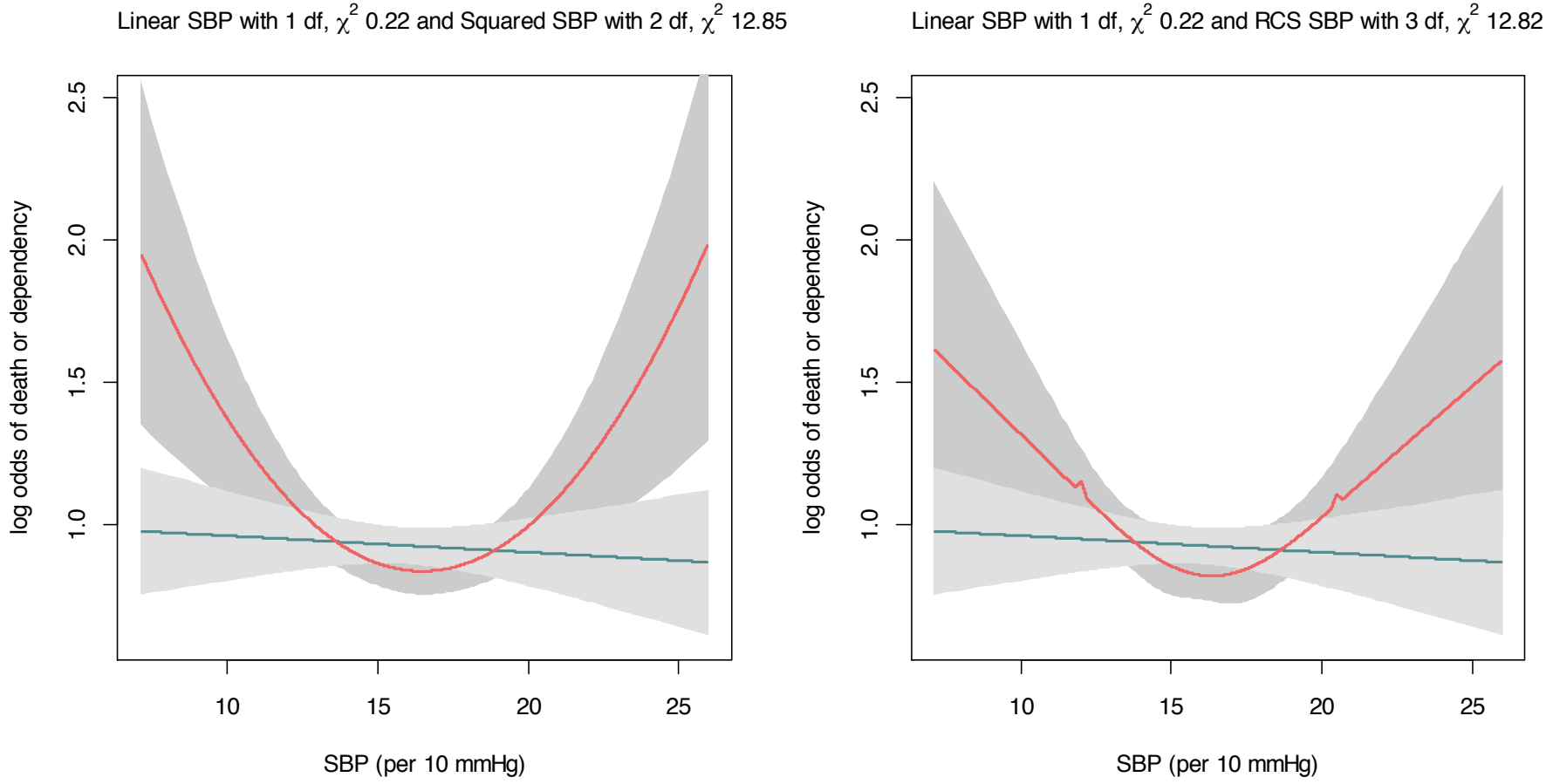


Figure 2-4 Example of a non-linear relationship. The left-hand plot describes a simple quadratic, whilst the right-hand side shows a 4 knot RCS fit

2.4.4.3 Assessing additivity

In equation (2.1) each term has one associated regression parameter. An interaction occurs when the presence of one predictor modifies that of another. The simplest example of this is for a continuous predictor X_1 and a binary predictor X_2, regressed upon Y. The function $g(\cdot)$ here represents some property, e.g., $E(Y \mid X)$, of the distribution of $Y \mid X$,

$$g(Y \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2.$$

When $X_2 = 1$, the categorical variable is present; the consequent impact on the slope of X_1 is given as,

$$g(Y \mid X_1, X_2 = 1) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1.$$

Evidence of an interaction between X_1 and X_2 therefore suggests that a different slope applies for X_1 dependent upon X_2. This is illustrated in the figure below (Figure 2-5). The additivity assumption will be assessed in this thesis by investigating the associated improvement in model fit when allowing two-way interactions (Harrell et al., 1996).

2.4.5 Model comparison

A formal method for selecting the best model fit (e.g., between a non-linear transformation and linear) is required. Given the observed data, the maximised likelihood (ML) is informative of the overall quality of the fit. Directly comparing the ML from two competing fits on the same data seems an appealing way to rank competing models. However, where a smaller model is contained within a larger model (i.e., nested), the model that fits more parameters will always have a larger ML over the smaller model. Two approaches for overcoming this issue are considered.

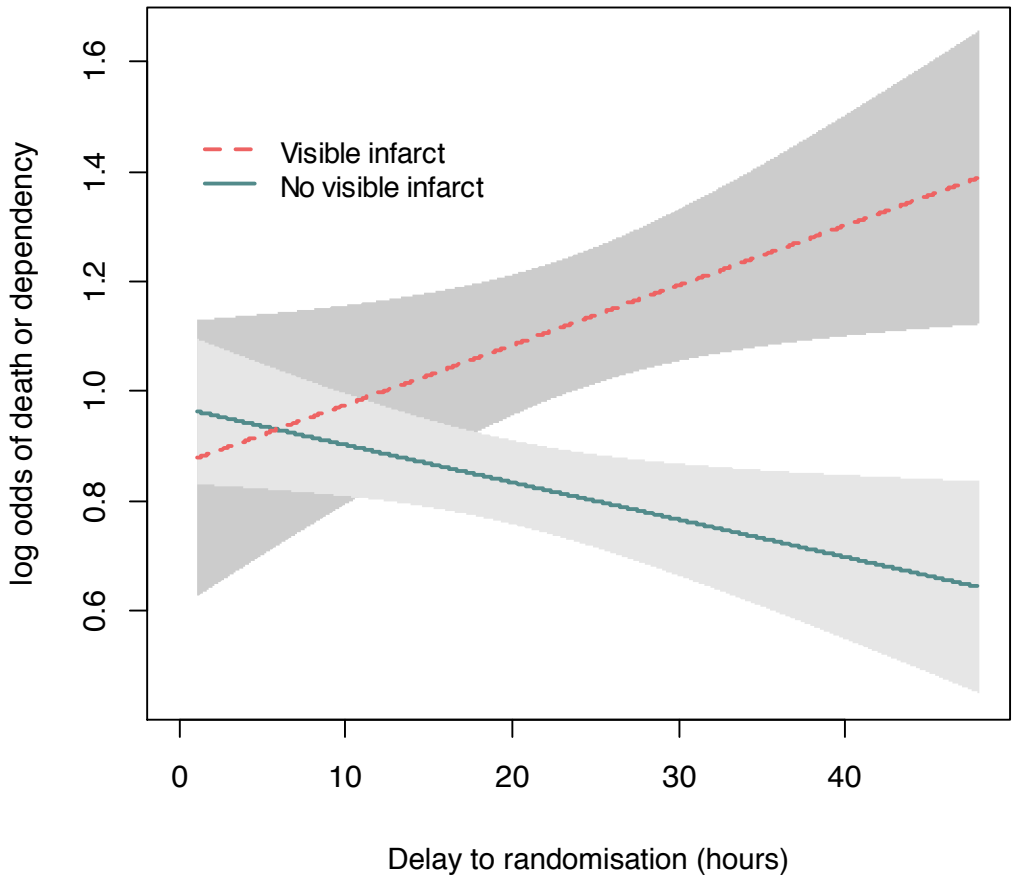


Figure 2-5 Example of an interaction

2.4.5.1 Penalised likelihood

Instead of using the ML directly, a penalty for the number of parameters fitted can be applied to account for the complexity of the fit relative to the improvement in ML (Christie et al., 2011).

$$l^* = \hat{l} - f(p,n)$$

Here $\hat{l}$ denotes the maximised log likelihood and $f(p, n)$ some penalty function based on the number of parameters modelled, p, and the size of the sample, n. One approach is to use Akaikes Information Criterion (*AIC*), which applies a penalty for the number of parameters only.

$$AIC = -2\hat{l} + 2p \qquad \textbf{(2.7)}$$

A reduction in *AIC* units indicates an improvement in the model fit. Another formulation of the *AIC* commonly encountered is the *adjusted* χ^2 which is given by the model $\chi^2 - 2p$, with p as specified before. Under this form an increase in *AIC* indicates an improved fit. A useful feature of the *AIC* is that there is no need for models to be nested provided they are fit using the same data.

Example From Figure 2-4 and using equation (2.7) the *AIC* for the simple linear fit of SBP (using 1*df*) to the log odds of death or dependency was 5383.7 and for the quadratic 5373.1 (using 2 *df*), an improvement in fit which translates as a reduction of 10.6 *AIC* units.

2.4.5.2 The likelihood ratio-test

A significance test can be used to compare two nested models where the null hypothesis is that the additional parameters have no additional value to the quality of the fit (Dobson, 1991).

$$-2(\hat{l}_0 - \hat{l}_1) \sim \chi^2_1$$

Example The log likelihood for the linear fit in Figure 2-4 was -2689.9 and -2683.5 for the quadratic fit. This gives a LR of 12.8 with a P-value of 0.0004, supporting the quadratic fit over the linear fit when compared to a $\chi^2_{(2-1)}$ where $\chi^2_{2-1,0.95} = 3.84$.

2.5 Model evaluation

The most important aspect of prediction is the performance of the model evaluated in new or existing data (Collins and Le Manach, 2013). This process is typically referred to as model *validation*; however the point will be made here that this is in fact a misnomer. The purpose is to *evaluate* the performance of the model; the notion of *validation* is somewhat pre-specified and confirmatory. In this thesis *model evaluation* will be used. Model evaluation is therefore informative of how well a given model meets its intended purpose, for example: *how well does a clinical prediction model predict the risk of a recurrent stroke*? There are two approaches denoted as *internal* evaluation and *external* evaluation (Altman et al., 2009).

Internal evaluation involves the partial or complete re-use of the original model development data. The most commonly adopted methods used to internally evaluate model performance are outlined (Steyerberg, 2009). The first and most appealingly simple is the *apparent* evaluation. Here model performance is assessed within the original data used in development. However, the model should be expected to perform well in these data due to overfitting; the resulting performance metrics are optimistically large. The only justification for assessing apparent evaluation is as a 'sanity-check' during the model development process to illustrate that the proposed fit holds some prognostic ability. Secondly, a *split-sample* evaluation may be used. This requires the investigator to employ a random split (e.g., a 50:50 split) of their obtained data which is then used to fulfil both the requirement of model development and of model evaluation. The issue with this approach is that the model is fitted only within a sub-section of the obtained data with a consequent impact upon the precision of the obtained estimates. This can lead to unstable effect estimates and additionally imprecise metrics of performance. The use of a random-split ensures that the data will represent a simple random-sample of the data with each side of the split having similarly distributed baseline features so that again performance will be

somewhat optimistic. The third method is *cross-validation*, which is similar to the split-sample approach only here the sample data are partitioned in such a way that a subsection of the dataset is left out of model development (e.g., one tenth) for the purpose of assessing performance. This is called a *K*-fold cross-validation process as it is repeated till each of the K^{th} subsections is used. One of the drawbacks of this approach is that it does not reflect all aspects of model uncertainty when variable selection methods are used suggesting more stability in variable selection than would be anticipated in practice. The final approach is *bootstrap* evaluation. Bootstrap evaluation is regarded as the most robust method for internally assessing model performance by *correcting for optimism*. Here the entire dataset is used in model development thus avoiding the small sample size problem associated with the split-sample approach. Model evaluation is undertaken in a series of samples drawn with-replacement from the original data of the same size as the original sample. For each draw the difference between metrics, i.e., apparent minus bootstrap replicate, is evaluated over all replications and used to estimate the optimism in the apparent performance. An average estimate of this optimism can then be subtracted from the original estimate providing an estimate of performance that is said to have been corrected for optimism.

Internal evaluation methods attempt to account for the fact that assessing model performance directly on the same data as used in development ultimately results in predictions that are over-fit to the data. It is not possible to assess the generalisability of the proposed model fit using internal evaluation and for this reason performance is best assessed on *external* data. External evaluation involves measuring model performance on a different set of patients recruited under different exclusion/inclusion criteria to that used in development, or perhaps within a different clinical setting (e.g., primary or secondary care). It is therefore important to understand how the development cohort compares to the evaluation cohort to fully judge the generalisability.

Regardless of whether the data source used is ‘internal’ or ‘external’ the metrics calculated are the same.

2.5.1 Overall model performance

Overall performance may be best measured by the amount of variance explained by the model, i.e., the *R*-squared (R^2). Although synonymous with linear regression, an analogue for binary outcomes exists in the form of Nagelkerke's R^2 logarithmic scoring rule (Nagelkerke, 1991).

$$\text{Nagelkerke's } R^2 = \frac{1-\left(L_0/L_1\right)^{2/n}}{1-L_0^{2/n}} \qquad \textbf{(2.8)}$$

Where n is the number of patients, L_0 is the likelihood for the model with no predictors (the null model) and L_1 is the model with predictors. This measure is scaled to range from 0 to 100%.

More generally, overall performance can be split into two constituent components: (i) discrimination; and (ii) calibration (Steyerberg et al., 2010).

2.5.2 Model discrimination

Discrimination summarises how well a model separates those with an event from those without. The concordance statistic or *c*-statistic is a rank order measure and is one way of measuring how well a given prediction model or risk score separates patients. For a given pair of patients, one with the event of interest and one without, the *c*-statistic is interpreted as the probability that a greater predicted risk (or higher risk score) is given to the patient with the event than the patient without. For a dichotomous outcome this corresponds to the area under the receiver operating characteristic curve (AUROCC). The Receiver Operating Characteristic (ROC) curve is a plot which maps *sensitivity* to *one minus specificity* for each potential cut-off of the score or predicted probability (see example Figure 2-6). The sensitivity of a test is the conditional probability of a positive result (i.e., a score or predicted probability above some fixed threshold) amongst events and the specificity of a test is the conditional probability of a negative result (i.e., below some fixed threshold) amongst non-events. The AUROCC can therefore be interpreted as the probability that a higher score or predicted probability is given to a patient with the event of interest over a patient without (Cook, 2007).

Let N_1 represent the number of patients with an event and N_2 represent the number without and $\hat{p}_{1,n_1}$ and $\hat{p}_{1,n_2}$ the respective individual predicted probabilities. The *c*-statistic is given by:

$$c = \frac{1}{N_1 N_2} \sum_{n_1=1}^{N_1} \sum_{n_2=1}^{N_2} \psi(\hat{p}_{1,n_1}, \hat{p}_{1,n_2}), \tag{2.9}$$

$$\text{where } \psi(\hat{p}_{1,n_1}, \hat{p}_{1,n_2}) = \begin{cases} 1 & \text{if } \hat{p}_{1,n_1} > \hat{p}_{1,n_2} \\ 0.5 & \text{if } \hat{p}_{1,n_1} = \hat{p}_{1,n_2} \\ 0 & \text{if } \hat{p}_{1,n_1} < \hat{p}_{1,n_2} \end{cases}$$

A non-parametric test for comparing two ROC curves was proposed by DeLong *et al*. which accounts for the correlation between the two vectors of predictions made by two separate models on the same data. A P-value can be obtained by comparing the resulting test statistic to the χ^2 distribution (DeLong et al., 1988).

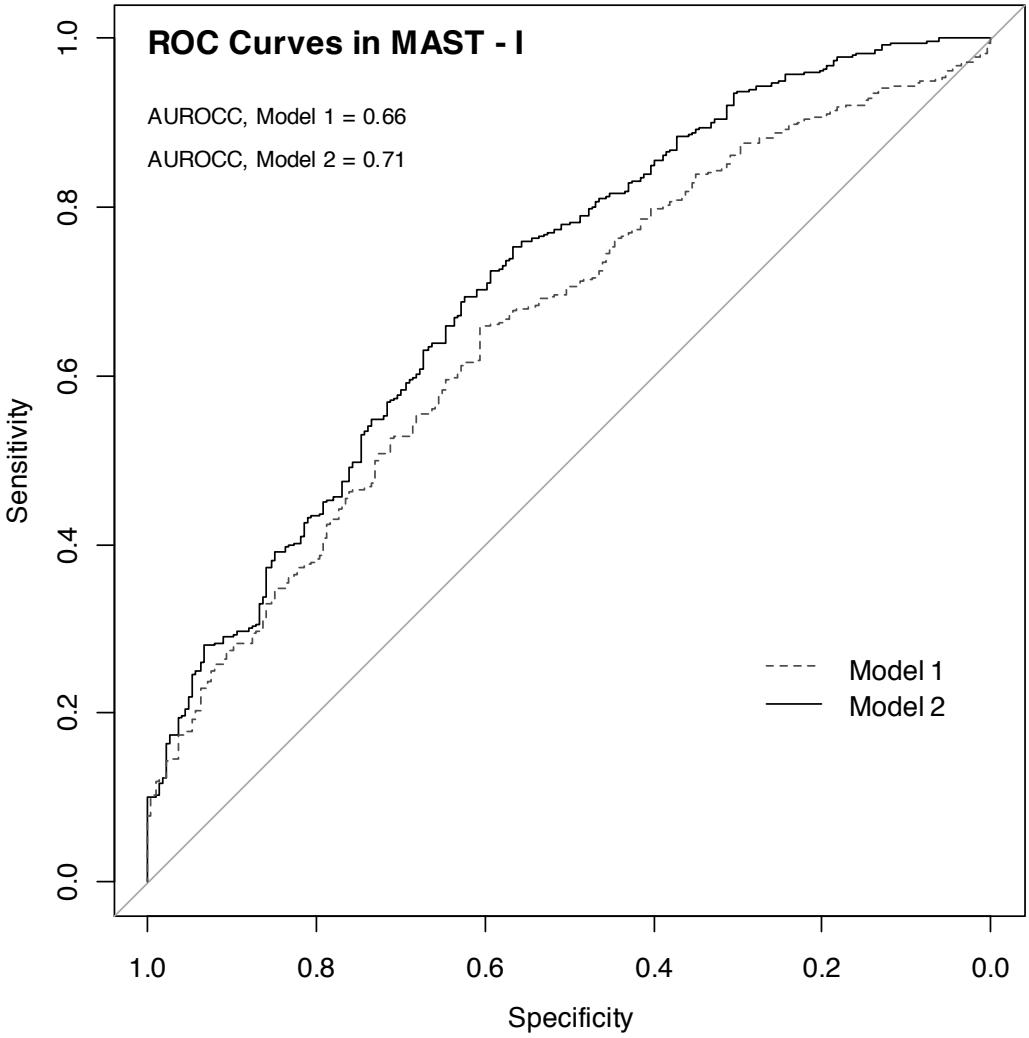


Figure 2-6 Example of a Receiver Operating Characteristic (ROC) Curve for two logistic regression models predicting six month death or disability. Model 1 (AUROCC = 0.66) contains a number of important predictors whilst Model 2 (AUROCC = 0.71) contains the same predictors with the addition of patients conscious state on admission

2.5.3 Model calibration

Calibration summarises how well the observed events match the predicted events. For example, if a model predicts the risk of recurrent stroke as 5%, then 5 out of 100 such patients should experience the event. The best representation of this is to plot grouped predicted risks along the *x*-axis (e.g., deciles of risk) against the observed fraction of events within each group on the *y*-axis (see example Figure 2-7). This contrasts the mean predicted risk to the mean observed frequency and is therefore a graphical display of the Hosmer-Lemeshow test of goodness of fit – a formal test which contrasts observed event rates to the expected event rates (Lemeshow and Hosmer, 1982). By estimating the slope and the intercept a better understanding of the bias in prediction can be obtained. A perfectly calibrated model has a slope of one and an intercept of zero. The slope indicates the degree of bias resulting from the model being: (i) over fit in development; and (ii) real differences in predictors. The intercept indicates differences in the prevalence of the outcome and the consequent systematic over (< 0) or under (> 0) prediction of risk. Estimates for these are outlined below.

2.5.3.1 Calibration-in-the-large

Calibration-in-the-large is the difference between the mean predicted risk and the mean of the new outcomes (i.e., in new data). For a non-linear model like a binary logistic regression, this difference must be evaluated on the log odds scale. However, the mean difference is not the same as the difference in the means. Therefore, when modelling this relationship some adjustment must be made for the corresponding intercept term (i.e., new predictions) achieved through an offset which enables a literal interpretation of the predicted risks (Steyerberg, 2009). The included offset term forces the coefficient for the linear predictor to equal one (Crowson et al., 2013).

2.5.3.2 Calibration slope

The calibration slope is calculated by fitting the linear predictor obtained through the application of an existing model to a new dataset in a logistic regression. Note that for both of these measures (slope and intercept) the corresponding standard errors allow confidence intervals and P-values from Wald tests to easily follow.

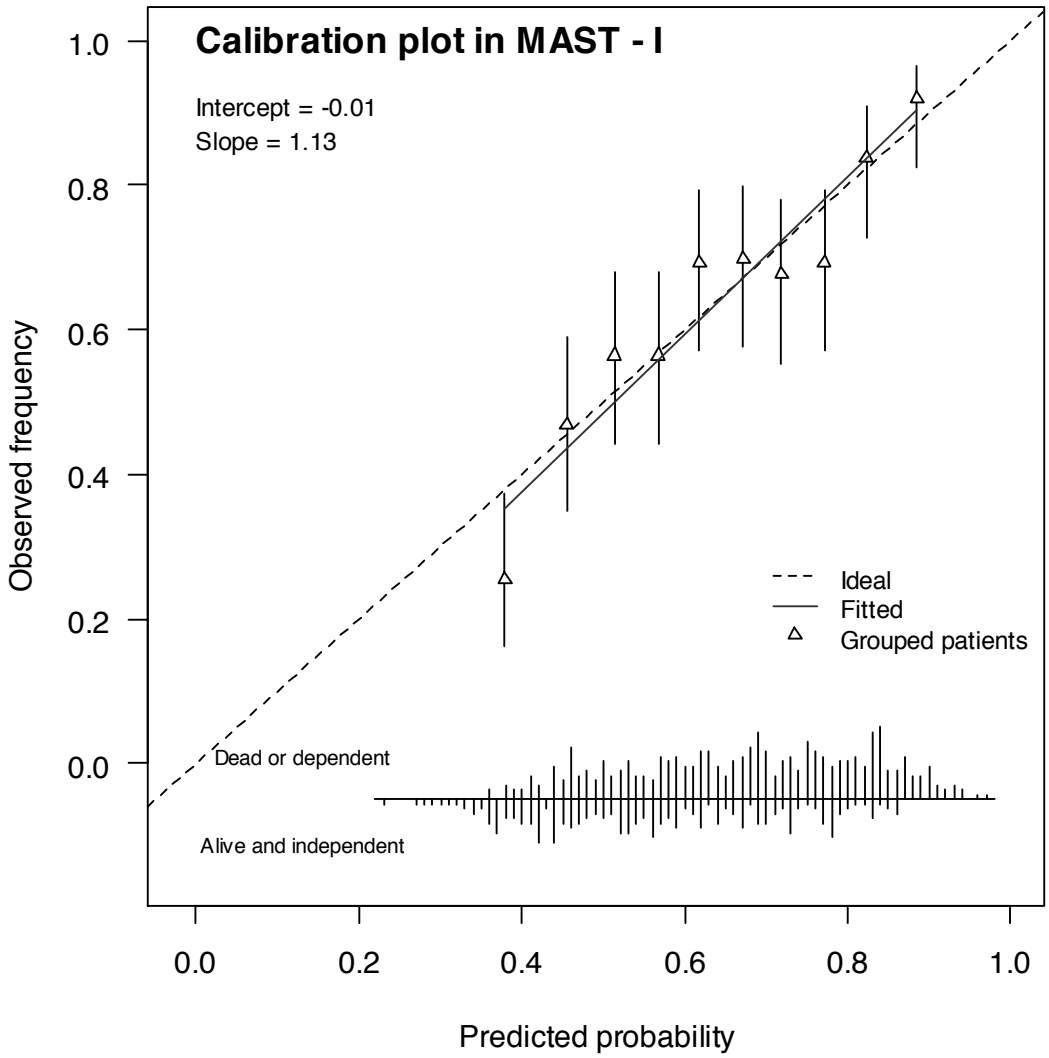


Figure 2-7 Example of a calibration plot. Each triangle denotes the observed frequency by grouped decile of predicted risk. Wilson 95% CI are provided about these estimates (Wilson, 1927).The distributions of predicted risk split by outcome are provided at the bottom of the plot. This is indicative of discrimination: the more distinct the two distributions are from one another the greater the discriminative ability of the model (AUROCC = 0.71)

2.5.4 The Net Reclassification Improvement

The AUROCC is difficult to interpret. Additionally, it has been highlighted that it is an insensitive measure for quantifying the added benefit of a new risk factor (Cook, 2007). Nancy Cook demonstrated this by considering a number of known predictors for the 10 year risk of CVD in healthy women, the same predictors in fact used in the established Framingham Risk Score. Interestingly, when low-density lipoprotein cholesterol (LDL-C) was included in a Cox model already containing age the *c*-statistic showed little incremental improvement moving from 0.70 to 0.71. This is despite the fact that LDL-C is known to be statistically significant and has a moderate effect size on the risk of 10 year CVD. Cook concluded that it is far more important to quantify the shift in predicted risks than to rely upon the *c*-statistic when assessing a new predictive marker. This paper prompted a bid to place a value on such a shift resulting in the Net Reclassification Improvement (*NRI*). The *NRI* uses reclassification tables to quantify the proportionate movement in events and non-events attributed to the inclusion of the new risk factor (Pencina et al., 2008). Consider two models: M_0, a base model and M_1, a model with all the predictors in M_0 but with one additional risk factor (e.g., a new biomarker or imaging variable). Two vectors of predicted probabilities, P_{M0} and P_{M1}, can be calculated on the same data. By categorising patients into groups of risk, separate cross classifications of events and non-events can be produced (as seen in Table 2-1). This is formalised below, closely following the exposition given by Pencina *et al*.,

$$NRI = P(up \mid Y = 1) - P(down \mid Y = 1) + P(down \mid Y = 0) - P(up \mid Y = 0) \qquad \textbf{(2.10)}$$

where $Y = 1$ signifies an event, and $Y = 0$ a non-event. Each of the conditional probability statements are estimated directly from the observed sample data as follows:

$$\hat{p}_{up,Y=1} = \hat{P}(up \mid Y = 1) = \frac{\#\,events\ moving\ up}{\#\,events}$$

$$\hat{p}_{down,Y=1} = \hat{P}(down \mid Y = 1) = \frac{\#\,events\ moving\ down}{\#\,events}$$

$$\hat{p}_{up,Y=0} = \hat{P}(up \mid Y = 0) = \frac{\#\,non\text{-}events\ moving\ up}{\#\,non\text{-}events}$$

$$\hat{p}_{down,Y=0} = \hat{P}(down \mid Y = 0) = \frac{\#\,non\text{-}events\ moving\ down}{\#\,non\text{-}events}$$

An estimate for the *NRI* is then obtained using each of these components.

$$nri = (\hat{p}_{up,Y=1} - \hat{p}_{down,Y=1}) - (\hat{p}_{up,Y=0} - \hat{p}_{down,Y=0}) \quad \textbf{(2.11)}$$

The *NRI* is the sum of four conditional probabilities and is therefore bound within the range $-2 \leq nri \leq 2$. An estimate of its standard error is given by:

$$SE(nri) = \sqrt{\frac{\hat{p}_{up,Y=1} + \hat{p}_{down,Y=1}}{\#\,events} + \frac{\hat{p}_{up,Y=0} + \hat{p}_{down,Y=0}}{\#\,non-events}} \quad \textbf{(2.12)}$$

This is used to give an asymptotic test of the null hypothesis that the *NRI* is equal to zero. Assuming independence between events and non-events, the associated *z*-statistic is given as the ratio of *nri* and its standard error.

Example Two arbitrarily defined categories are used to categorise the predicted risks obtained from two logistic regressions for the risk of poor outcome at low and high risk defined as ≤50% and >50% respectively. A re-classification table is provided in Table 2-1. Using equation (2.11) and equation (2.12) the *NRI* and standard error were estimated as 0.0701 and 0.0275 respectively. Under the assumption of asymptotic normality the 95% CI for the *NRI* estimate was 0.0162 to 0.1240, and the associated P-value was 0.0108.

Table 2-1 Example of reclassification for the predicted probabilities with and without patients conscious state on admission amongst events and non-events in MAST-I

Model without consciousness	Model with consciousness		
Frequency (Row %)	≤ 50%	> 50 %	Total, %
Participants who experience an event			
≤ 50%	35 (54)	**30 (46)**	65 (16)
> 50 %	18 (5)	313 (95)	331 (84)
Total, %	53 (13)	343 (87)	396
Participants who do not experience an event			
≤ 50%	72 (91)	7 (9)	79 (35)
> 50 %	**16 (11)**	131 (89)	147 (65)
Total, %	88 (39)	138 (61)	226

Note: dark grey shaded area indicates movement in the correct direction whilst light shaded area indicates movement in the wrong direction

A myriad of publications and citations arose in the wake of the Pencina *et al*. article with the *NRI* rapidly adopted amongst medical journals and fiercely critiqued within statistical journals (Vickers and Pepe, 2014). A recent review from Kerr *et al*. summarised the difficulties with the *NRI* drawing particular attention to its many misinterpretations and weaknesses (Kerr et al., 2014). The main drawbacks are provided here. First, the *NRI* is commonly quoted as a percentage; however (as mentioned above), the *NRI* combines four conditional probabilities and is therefore bound within the range [-2 to 2]. Second, combining the component reclassification measures (i.e., events and non-events) is of little benefit and in actual fact limits the interpretation considerably. Explicitly stating these unique values alongside their precision and reclassification tables is far more informative. Third, for three or more categories the *NRI* gives equal weight to all movement, despite the fact that a shift from high risk to medium risk may have a comparably smaller impact in the decision making process than a move from high to low risk. Finally, determining how the obtained predicted risks should be categorised in the absence of any pre-specified groups by necessity will be arbitrary.

Indeed, Pencina *et al*. have extended their *NRI* to a *category-free*, or so called *continuous NRI* (Pencina et al., 2011). However, Kerr *et al*. also highlight a number of issues with this measure, including: (i) uninformative new markers yielding beneficial *NRI*s; (ii) small and clinically irrelevant changes in the predicted risk have an equal contribution in the calculation; and (iii) its interpretation is undefined.

Importantly, Kerr *et al*. conclude that the NRI indices should not be used in the context of hypothesis testing especially when far superior tests already exist for this purpose, for example, the likelihood ratio test (see section 2.4.5.2).

2.6 Missing data

Missing data arise frequently in practice. Appropriate actions must be taken when handling such data – these actions should be explicitly outlined and justified.

2.6.1 The mechanisms of missingness

In essence missing data represent gaps in the available data: entries that were lost or else never recorded during recruitment to the study. Missing data act through three distinct mechanisms (Steyerberg, 2009). First, data may be missing by accident, e.g., an accident in the lab which resulted in a set of blood samples being destroyed (Missing Completely At Random – MCAR), in which case missingness is independent from the data observed. Second, for a given variable entries may be missing as a direct result of other measured variables (Missing At Random – MAR) in which case a dependency exists between the probability of missingness and some or all of the observed covariates. Finally, data can go missing for some hidden reason (Missing Not At Random – MNAR), i.e., the probability of a missing value is dependent upon a measure which was never recorded or may even depend on the missing values themselves.

2.6.2 Handling missingness

There are various methods for handling missing data. The most commonly encountered method is the complete case analysis. Here subjects with missing observations are removed from the dataset and hence removed from the formal analysis. Should an investigator decide to report both univariate and multivariable results then it becomes important that explicit references be made to which sample-size each analysis corresponds to (Harrell, 2001). The complete case approach incurs a reduction in the sample size and implicitly assumes that any missing values are MCAR. A consequence of this is a reduction in the statistical power, and, where the MCAR assumption fails, the production of biased effect estimates (Vergouwe et al., 2010). Imputation methods are used as a way of sensibly 'filling in the gaps' so that all measurements on all subjects contribute to the analysis. This is typically internal and will be based on some or all of the measured characteristics and their inter-dependencies. There are numerous methods available. The simplest approach for

continuous variables is to replace any missing entries with the mean of that variable, whilst for categorical variables missing entries are replaced with the most frequently observed level. The main drawback of this approach is that it ignores additional information regarding the correlation structure between the measured variables. A more sophisticated approach utilises this information by regressing the variable with missingness upon the additional observed characteristics producing a statistical model specified for the explicit prediction of missing entries. This method requires that the missing entries have arisen at least in part through the MAR mechanism. A single draw is then made generating one complete dataset which is then analysed. However, this does not truly reflect the uncertainty that is present in the imputation process since on a second draw a slightly different dataset is obtained which once analysed will yield slightly different results. This issue is mitigated by generating multiple sets of imputed data each of which is then analysed and combined using Rubin's rules (Rubin, 1996).

2.6.3 The imputation model

Provided that the missing data are indeed MAR (at least in part) then the imputation model will act as the link between those values that are missing and those that have been observed. The form taken in prediction is dependent upon the type of variable to be imputed, for example a binary variable would imply that a logistic regression be used, whilst a continuous variable suggests a normal linear model. The variables to be included are of three kinds: (i) those variables of interest to the analysis proper, e.g., the covariates to be included in a clinical prediction model; (ii) those variables that are of importance in the prediction of *missingness*, and regarded as auxiliary; and (iii) the outcome. Including the outcome in the imputation model is often thought to be somewhat *self-fulfilling*, however, if omitted, then covariate values are imputed as if not related to the outcome which can lead to biased results (Moons et al., 2006).

The imputation model takes the k^{th} variable of the K available variables and uses the remaining k-1 variables to predict those missing entries in k. This process is repeated for each of the K variables with missing entries. It is generally the case that in practice there will be missing values amongst the k-1 variables used in the imputation

of missingness within *k*. This is handled by using a *chained equations* approach (Azur et al., 2011). First some value must be used to create some *pseudo* complete data; a *first use* point is defined, e.g., a single imputation using the mean. This is a temporary imputation though which will allow an imputation model to be devised for each variable. Each prediction model is then used to impute the originally missing values thus replacing the *first use* mean values used at the beginning with a predicted imputation. This process is repeated for a pre-specified number of iterations until stability is achieved. This then provides the first of the *m* complete datasets. A further complication arises when handling missing outcomes in imputation. Provided that MAR holds for the missing outcomes, there is a small gain in efficiency to be made by including cases with missing outcomes in the prediction of missing covariate information (Little, 1992). However, the log likelihood for cases with imputed outcome values is equal to zero, so retaining imputed outcomes equates to adding noise to estimates (Von Hippel, 2007).

2.6.4 Rubin's rules

Having obtained *m* datasets and *m* associated analyses, pooled estimates are generated using Rubin's rules (Rubin, 1996). Consider some parameter, Q, and its corresponding variance, U, across m multiple sets two vectors are obtained: $\hat{\boldsymbol{Q}} = [\hat{Q}_1,...,\hat{Q}_m]$ and $\hat{\boldsymbol{U}} = [\hat{U}_1,...,\hat{U}_m]$, where $\hat{U}_i = Var(\hat{Q}_i)$. Rubin suggested a method for combining multiple estimates and calculating an associated standard error for the pooled estimate. These steps are summarised as follows:

1. Calculate the mean of the *m* estimates $\quad \bar{\hat{Q}} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i$

2. Calculate the within imputation variance $\quad \bar{\hat{U}} = \frac{1}{m}\sum_{i=1}^{m}\hat{U}_i$

3. Calculate the between imputation variance $\quad B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \bar{\hat{Q}})^2$

4. Estimate the total variance associated with $\bar{\hat{Q}}$ $\quad T = \bar{\hat{U}} + \left(1 + \frac{1}{m}\right)B$

2.7 Discussion

Dawid and Senn elegantly describe the philosophical issues that relate to the practice of constructing models from data (Dawid and Senn, 2011). They begin by noting that by their very nature models are abstract. They highlight the importance of refraining from 'reification', that to treat the model as if it truly exists is erroneous. A model will not, nor should it be expected to, exactly fit reality. By directly comparing models a better description of reality may be reached, though there are those levels of uncertainty that will always persist: (i) *stochastic* uncertainty, an intrinsic part of a probability model; (ii) *statistical* uncertainty relating to the limitations of knowledge given the availability of data; (iii) *model* uncertainty regarding the limits of knowledge about how the world works; (iv) *philosophical* uncertainty relating to the two schools of thought in statistics (i.e., frequentist or Bayesian); and finally (v) the unknown unknowns.

The following quote from Box and Draper is often used to summarise the issues associated with the modelling of data (Box and Draper, 1987):

> "Essentially, all models are wrong, but some are useful."

Indeed, prediction models can be *incredibly* useful when used in the correct context and setting. More recently, Thaddeus Tarpey in his presentation to the Joint Statistical Meetings in 2009 emphasised that models are not wrong, but that treating them as the truth is (i.e., the same *reification* fallacy noted by Dawid and Senn). He offered a more positive way of framing the Box and Draper quote (Gelman, 2012):

> "All models are right…most are useless"

Chapter 3: Predicting recurrent stroke and MI after stroke

Background and summary

This chapter presents a systematic review of the development of clinical prediction models for recurrent events in acute ischaemic stroke. A meta-analysis of model evaluation studies is followed in order to describe model performance. It is found that those currently available models appear to discriminate only modestly between patients with and without recurrent events: this may be explained at least in part by design weaknesses.

3.1 Introduction

Recovery after stroke can be complicated by recurrent events, e.g., stroke, myocardial infarction, and gastro-intestinal bleeding. The risk of recurrence is around 9% two years following the initial stroke (Azarpazhooh et al., 2008, Burn et al., 1994). Suppose a deterministic rule existed whereby it was *a priori* possible to say whether a given patient *would* or *would not* experience a complication. It logically follows that such a rule could influence the course of therapy. Whilst a perfectly discriminating rule (i.e., with an AUROCC of 1) is seldom available, a multivariable combination of measurements taken from patients on admission can often yield a sensible estimate of the underlying risk which can be of some value (Hayward et al., 2006).

Multivariable clinical prediction models have the potential to better inform both doctor and patient about patient prognosis by providing an absolute estimate of individual patient risk. The clinical utility of the model depends on a number of factors; including its perceived discriminative ability. For instance, a model developed to predict the risk of recurrent ischaemic stroke or MI for recent ischaemic stroke patients will describe the underlying heterogeneity for an observed group of

patients: some of whom will have a low predicted risk; and some at a high predicted risk. Provided that those with events in follow-up are given higher predicted risks than those free from events in follow-up then the prediction model may be useful in practice. It is well understood that better knowledge on patient prognosis has the potential to influence many facets of health care (Hemingway et al., 2013). In particular, informing a patient of his or hers predicted prognosis can play an important role in making treatment decisions (Moons et al., 2012b). This personalised approach to medicine places informed decisions in the hands of the patient and can be especially useful when emphasising those modifiable risk factors. However, models for recurrent events after stroke are currently not in widespread use. It is not clear why this is the case. Systematic reviews of clinical prediction models in other conditions (e.g., type II diabetes, cancer, traumatic brain injury) have identified a common set of flaws which, but to name a few, include: the use of poor methods; the inappropriate handling of missing data and the arbitrary categorisation of continuous variables (Collins et al., 2011, Mallett et al., 2010, Altman, 2009, Perel et al., 2006).

This chapter presents a systematic review of those existing models for the prediction of recurrent stroke or MI after stroke. There are two broad aims: (i) to explore the quality of the cohorts and statistical methodology used in model development; and (ii) to summarise any evidence of their external performance.

3.2 Methods

An analysis protocol was developed (Thompson, Murray and Whiteley) and circulated amongst the convenors of the Cochrane Prognosis Methods Group (http://prognosismethods.cochrane.org/welcome) for comment. This protocol was made available online prior to conducting the electronic search: http://www.dcn.ed.ac.uk/dcn/documents/profile_protocols/whiteley_p2.pdf.

Medline and EMBASE databases were searched from 1980 to the 19th of April 2013 with an electronic search strategy using a search term for ‘stroke’ and synonyms for ‘clinical prediction models’ (see Appendix A Table 3-4). Additional checks for any missed articles were made using article reference lists, personal files and ‘Google

Scholar' (http://scholar.google.com/) for citations of relevant articles. This review adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist (see http://www.prisma-statement.org).

3.2.1 Inclusion criteria

Eligible articles developed and/or evaluated a multivariable clinical prediction model for the risk of recurrent ischaemic stroke, myocardial infarction or all vaso-occlusive arterial events in cohorts of adult patients with ischaemic stroke (or mixed ischaemic stroke and TIA). Articles had to provide their model in sufficient detail such that predictions could be made for new patients. An inception date of 1980 was used to account for the introduction and routine use of CT scans in clinical practice. Risk factor studies presenting a univariable model were excluded. Any studies using cohorts that included haemorrhagic stroke patients at baseline or any haemorrhagic events in follow-up were excluded. No language restrictions were made.

3.2.2 Data extraction

One author screened all titles and abstracts identified by the electronic search against the inclusion criteria prior to full text assessment (Thompson). Two authors (Thompson and Whiteley) extracted data independently with a detailed data extraction form that was developed and piloted by three of the authors (Thompson, Murray, and Whiteley). Discrepancies were resolved by discussion.

No recommended tool for appraising the quality of prediction models currently exists. Quality items were therefore adapted from similar systematic reviews to ensure that all qualitative assessments made in this review were in-keeping with previous such studies (Counsell and Dennis, 2001, Hayden et al., 2006, Mallett et al., 2010, Mushkudiani et al., 2008, Collins et al., 2011, Kwakkel et al., 1996, Laupacis et al., 1997). These items are discussed in detail below.

Two types of articles were distinguished: (i) development studies reporting the construction of a prediction model, and (ii) evaluation studies (also known as validation studies) assessing model performance in a cohort of new patients.

3.2.3 Qualitative assessment of development studies

The quality of a development study may be best assessed through its: (i) *reporting*; and (ii) *methods*. It is evident that the first aspect overlaps the second since a poorly reported study could detract from, or obscure, the methods used. It is argued that authors must be transparent and meet a high standard of reporting thus enabling consumers of the research to weigh up the value that the prediction model offers in influencing clinical practice. In principle, the process should be reported in enough detail that the model development could be independently reproduced, put simply: the same rigor expected in the reporting of RCTs should be adopted in model development studies (Groves and Godlee, 2012).

3.2.3.1 Internal validity

The sample of data used in model development is integral to the quality and generalisability of the proposed fit. The data should be viewed as the foundations of the model: if these foundations are weak or inappropriate for answering the research question then any inferences will very likely be subject to various biases. Prospectively collected data are typically of greater quality than retrospectively collected data and are therefore preferred in model development (Moons et al., 2009).

Loss to follow up is common. Authors should state the number of patients lost along with any explanations for these losses. An arbitrary proportion often thought adequate for analysis is anything between 80 to 95% complete follow-up (Fewtrell et al., 2008). The completeness of follow-up, which takes into account the duration of follow-up can be an informative way of reporting follow-up in survival data (Clark et al., 2002).

Predictor variables, outcomes, and the time window across which they apply should be explicitly defined and reported. Predictions obtained using ill-defined variables can be difficult to generalise and in some instances may even be invalid.

An exhaustive summary of the extent of missing data and how such data were handled must be provided. Complete-Case (CC) analyses should be avoided in

favour of Multiple Imputation (MI) methods (Rubin, 1996, Vergouwe et al., 2010). As a general rule of thumb it has been suggested that imputation should be considered where the proportion of missingness exceeds 5% of the data. The impact of bias caused by proportions beneath this level may be negligible though still warrant investigation (Harrell, 2001, Schafer, 1999).

3.2.3.2 Statistical validity

Authors should avoid using data dependent predictor selection methods (e.g., stepwise selection) where possible (Sun et al., 1996, Steyerberg et al., 1999). As discussed in Chapter 2, these methods have various shortcomings which make them only suitable when there is a lack in background clinical knowledge. Clinical knowledge should always be used to inform the selection of risk factors. These risk factors should be retained within the final prediction model irrespective of their statistical significance (i.e., forced inclusion).

Continuous variables should be kept continuous and modelled as such. Arbitrary categorisations are often adopted without justification and should be avoided. Dichotomisation is both statistically and clinically inefficient (Altman and Royston, 2006). It equates to a needless reduction in statistical power as well as treating patients close to the respective cut-points as if entirely different. Such steps in risk are of course biologically implausible. The figure below illustrates the problems associated with categorising continuous variables by contrasting a simple continuous linear fit to a binary categorical fit (Figure 3-1). The relationship between the predictor variable X (e.g., patient age) and the binary outcome Y (e.g., death) on the log odds scale is assumed linear. A patients risk from the outcome therefore depends upon their value for X. Alternatively, the categorical fit assumes a constant risk for all of those where $X<70$ and a single step increase for those where $X\geq70$. Altman and Royston stressed that if cut-points are used then they should be based on clinical reasoning, blinded to the observed data and ideally defined across more than two levels (Altman and Royston, 2006). Optimal cut-points based on P-values should never be used.

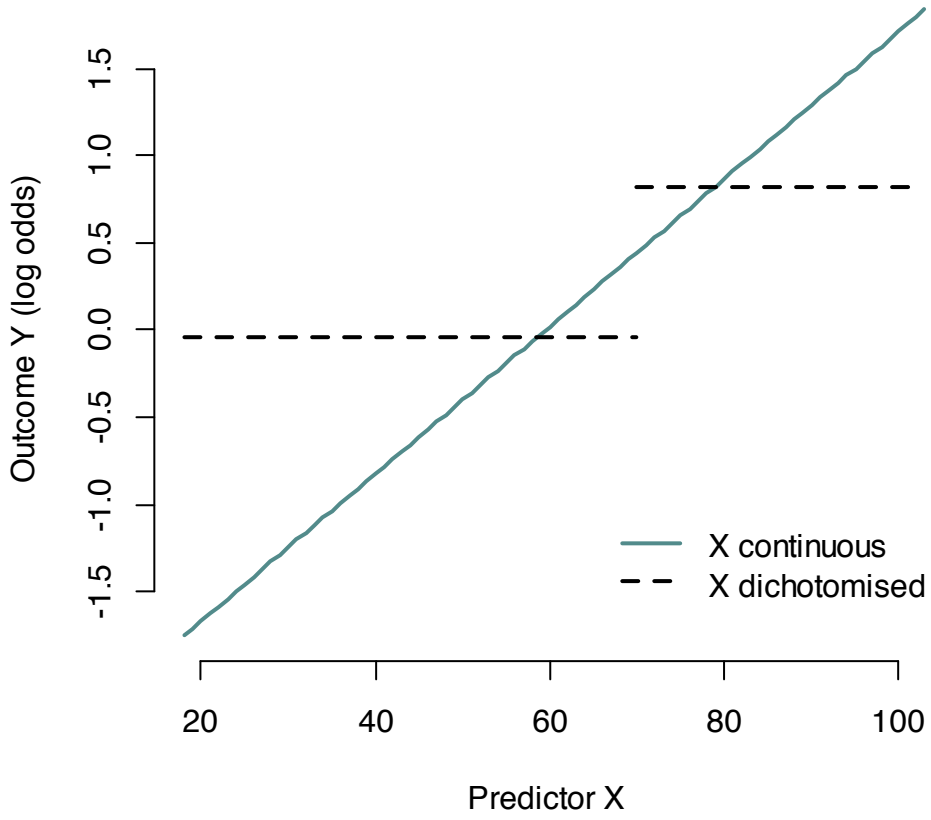


Figure 3-1 Example plot demonstrating the relationship between a continuous predictor, X, and the log odds of outcome for: (i) linear X; and (ii) X dichotomised at 70.

The sample size available for model development (hereafter referred to as the derivation sample) must be reported along with a sufficient description of the baseline characteristics, missing data and the exclusion of any subgroups. The number of patients experiencing an event in follow-up (hereafter referred to as the effective sample size) should be stated explicitly. This has great importance in evaluating how over-fit the model may be relative to the selection of covariates considered. Peduzzi *et al*. demonstrated through simulation that a minimum of 10 to 15 events per fitted parameter should be used as a guide for sensible parameter estimation: anything below this risked considerable biases in parameter estimates (Peduzzi et al., 1996). This rule of thumb can be used as an upper limit for the number of predictors that can be introduced to a model. Note that the *effective sample size* for a binary outcome is the smaller of the two numbers in the ratio of events to non-events. As discussed above, selecting covariates for inclusion in a

clinical prediction model should ideally be based on pre-existing clinical knowledge. Data dependent predictor selection, e.g., stepwise selection or univariate screening against some pre-specified significance level (e.g., P-value < 0.05) have many limitations, one of which is that they lead to an overfitting of the model by capitalising on data specific quirks and idiosyncrasies (Harrell, 2001). Michael Babyak gives an excellent overview of the points discussed above regarding over-fit models, describing a strategy on how to minimise the risks of overfitting through cautious expenditure of the degrees of freedom and shrinkage methods to correct over-fit regression coefficients (Babyak, 2004).

3.2.3.3 Evaluation of the model

The performance of a prediction model is best assessed within a new cohort of patients. Authors frequently state the performance of their model for the same data used for development (training data). This is equivalent to a student sitting last year's paper as this year's exam: the student training on past papers will perform well but may be thrown by a set of unseen questions. External validation in new data (test data) offers the most rigorous assessment of model performance and generalisability. A description of the baseline characteristics in the validation cohort should be reported to enable a comparison of the *validation cohort* to the *development cohort* suggesting how well the model may perform and ultimately summarising how the model may be generalisable (e.g., amongst older/younger patients etc.). Model performance must be assessed quantitatively via discrimination (e.g., the Area Under the Receiver Operating Characteristic Curve (AUROCC)) and calibration metrics (e.g., Hosmer-Lemeshow test or calibration plots) (Steyerberg et al., 2010). Although external validation in new data is preferred, so called internal validation techniques (e.g., bootstrap re-sampling or *k*-fold cross-validation) can also be applied. These methods re-use the development data to some degree, adjusting the apparent performance (that observed in the training data) for optimism. Internal validation, however, is only as useful as the quality of the data used in model development, for example, no adjustments can be made to correct for centre specific attributes when using data from a single hospital cohort.

It is noted that the phrase 'model *validation*' is somewhat self-fulfilling. This unfortunate terminology persists in the literature but will be referred to as 'model *evaluation*' throughout this thesis.

3.2.4 A brief overview of methods in meta-analysis

In this section an overview of meta-analysis is provided covering the basic concepts regarding: the pooling of information; and the assessment of heterogeneity.

Often a number of studies assessing the efficacy of a given treatment with similar trial protocols will have been conducted. It would be convenient to summarise this information as a single 'pooled measure'. Meta-analysis is the branch of statistics concerned with the quantitative synthesis of such information where the goal is to provide a single, sensible, pooled measure of treatment effect (Deeks et al., 2008). There are two commonly adopted frequentist approaches in meta-analyses (Borenstein et al., 2010). A *fixed effect meta-analysis* assumes that each study measures the same true and fixed effect; therefore any observed variation is due to random error. Alternatively, a *random effects meta-analysis* says that estimates will vary from study to study due to random error and also due to true variation in the effect size assuming that the individual study estimates, θ_i, come from some distribution, e.g., $\theta_i \sim N(\theta,\tau^2)$ with variance,

$$\hat{\tau}^2 = \frac{Q-(k-1)}{\sum_{i=1}^{k} w_i - \left(\sum_{i=1}^{k} w_i^2 \Big/ \sum_{i=1}^{k} w_i\right)} \qquad \textbf{(3.1)}$$

where

$$Q = \sum_{i=1}^{k} w_i \left(\theta_i - \sum_{i=1}^{k} w_i\theta_i \Big/ \sum_{i=1}^{k} w_i\right)^2,$$

and

$$w_i = \frac{1}{SE(\hat{\theta}_i)^2}.$$

Here, Q is the heterogeneity statistic describing the sum of squared differences of the individual effect estimates from their weighted mean (N.B., this statistic has a Chi-

squared distribution and can be used to provide a test of heterogeneity with $k - 1$ degrees of freedom). The weights, w_i, denote the i^{th} inverse variance weight associated to each of the k studies: imprecise estimates (i.e., large SE) are given less weight whilst precise estimates (i.e., small SE) are given more weight. To account for the degree of variability due to real study differences in the pooled estimate, DerSimonian and Laird (DerSimonian and Laird, 1986) proposed using,

$$\bar{\theta}_{DL} = \sum_{i=1}^{k} w_i' \theta_i \Big/ \sum_{i=1}^{k} w_i' \tag{3.2}$$

with weights

$$w_i' = \frac{1}{SE(\theta_i)^2 + \tau^2}.$$

The precision of the estimator, $\bar{\theta}_{DL}$ is given by,

$$SE(\bar{\theta}_{DL}) = \sqrt{\frac{1}{\sum_{i=1}^{k} w_i'}}, \tag{3.3}$$

which can be used to produce a 95% CI, $\bar{\theta}_{DL} \pm z_{1-\alpha/2} SE(\bar{\theta}_{DL})$.

When reporting the summary estimate from a random effects meta-analysis the associated prediction interval (PI) should also be provided (Higgins et al., 2009). The 95% PI is interpreted as a plausible range within which an unknown estimate will be expected to lie in 95% of future samples (Riley et al., 2011).

$$\bar{\theta}_{DL} \pm t_{k-2;1-\alpha/2} \sqrt{\hat{\tau}^2 + SE(\bar{\theta}_{DL})^2} \tag{3.4}$$

Where $t_{k-2;1-\alpha/2}$ is the $100(1 - \alpha/2)\%$ from a t-distribution with $k - 2$ degrees of freedom.

3.2.4.1 The meta-analysis of AUROCC values

Area Under the Receiver Operating Characteristic Curve (AUROOC) obtained from evaluation studies were meta-analysed in a similar way to that adopted by Meads *et al* in their review of breast cancer incidence models (Meads et al., 2012). The idea of

using inverse variance weights was discussed above. Frequently though the standard error (*SE*) of the AUROCC estimate will not be provided, instead, only the 95% CI will have been reported. An estimate for the *SE* can be obtained by rearranging the approximate confidence interval: $AUROCC \pm z_{1-\alpha/2} SE(AUROCC)$. With a confidence interval ranging from [*a* to *b*] the *SE* can be estimated as,

$$\frac{b-a}{2(z_{1-\alpha/2})}.$$

In almost all situations the CI reported is a 95% CI and therefore the critical value for *z* is 1.96 giving a denominator of 3.92. A loss of accuracy can be expected as CIs for AUROCC estimates are typically only reported to two decimal places.

Individual evaluation studies will yield different model performance metrics as according to differences by chance and differences attributable to study design, i.e., patient populations, selection etc. It is for this reason that a random effect meta-analysis model will be used to summarise model performance. If three or more studies assessed the same model's performance within separate patient populations, then a random-effects meta-analysis was undertaken (DerSimonian and Laird, 1986, Viechtbauer, 2010). All measures of model performance were extracted along with any associated measures of uncertainty (e.g., 95% confidence intervals (CI) or standard error). Publication bias was assessed using funnel plots.

3.3 Results

A total of 12,456 articles were screened by title and abstract (Figure 3-2), thirteen of which were eligible for review. A further ten were identified from reference list checks and forward citation searches in Google Scholar. This comprised twelve development studies (Table 3-1) developing a total of 31 models (a median of 2 per study, interquartile range (IQR) 1 to 3) and eleven evaluation studies evaluating the performance of four models (Table 3-2). One relevant study written in a language other than English was included which was translated by Miss Cristina Matthews of the Centre for Population Health Sciences, University of Edinburgh (Alvarez-Sabin et al., 2008).

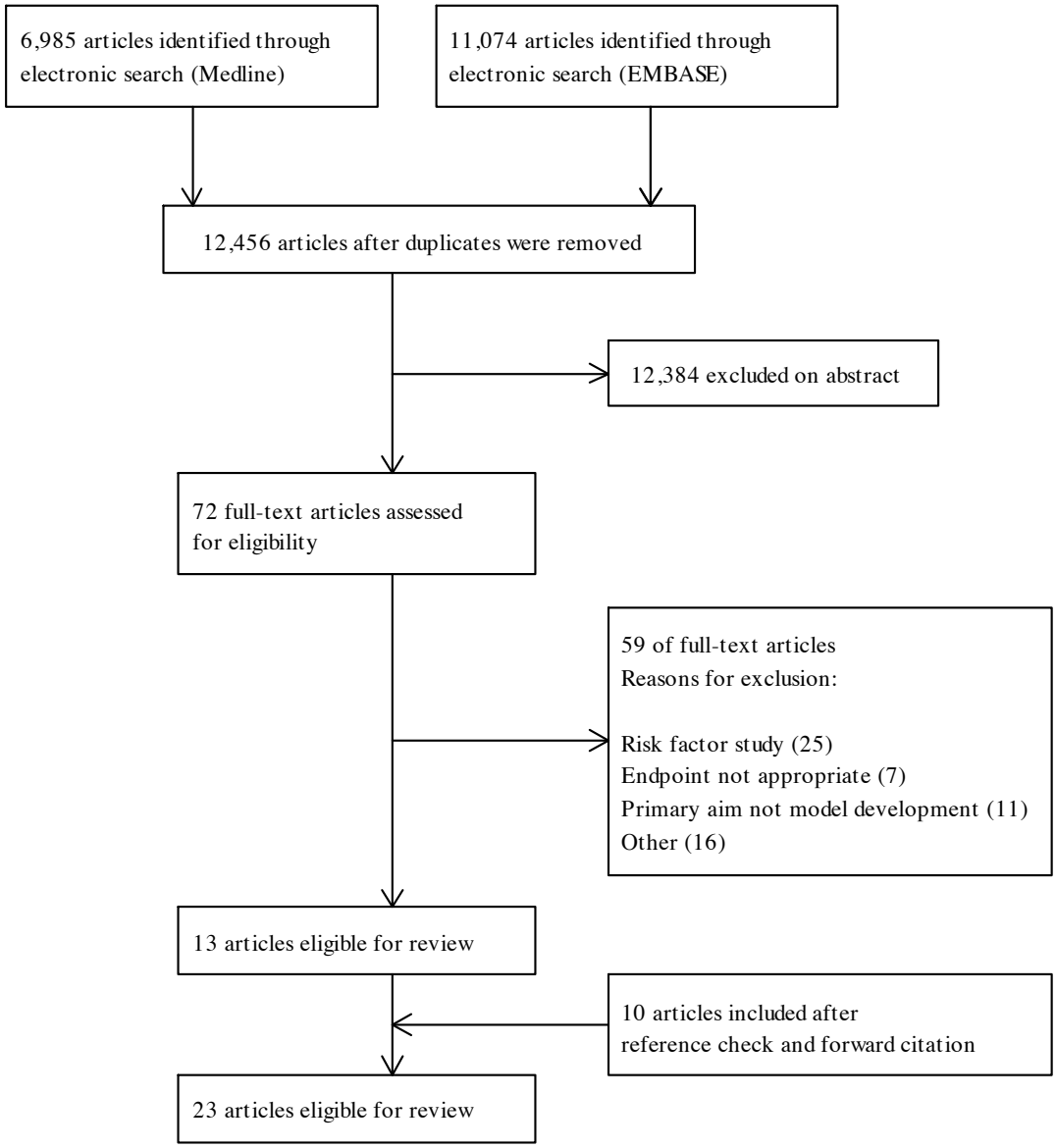


Figure 3-2 PRISMA flow diagram of selected studies

Table 3-1 Characteristics of 12 development studies presenting prognostic models for recurrent vascular events after ischaemic stroke

Author (model)	Year	Study design	n/N	Baseline event	Outcome	Time of outcome
Ay et al (RRE90)	2010	Retrospective single centre	60/1458	IS < 72 hours	IS	≤90 days
Dhamoon et al	2007	Prospective population cohort	102/655	IS	MI; or vascular death	5 years
Diener et al (ESRS)	2005	Prospective RCT	NA/6431	1 week ≤ IS ≤6 months	Stroke recurrence	1 year
Kamouchi et al (FSRJ)	2012	Prospective multicentre and retrospective registry	175/3067	IS ≤ 7 days	IS	1 year
Kernan et al (SPI-I)	1991	Retrospective single centre	38/142	TIA or minor stroke	Stroke or death	2 years
Kernan et al (SPI-II)	2000	Prospective RCT	90/525	TIA or minor stroke <90days	Stroke or death	2 years
Pezzini et al	2009	Prospective multicentre	73/511	First ever IS aged 18 to 45	Ischaemic recurrence: fatal/nonfatal MI, IS, or TIA	4 years
Putaala et al	2010	Prospective single centre	72/807	First ever IS aged 15 to 49	Fatal/nonfatal IS; or fatal/nonfatal stroke; or MI	5 years
Stahrenberg et al	2013	Prospective cohort	23/197	IS	Cardiovascular events	1 year
Sumi et al	2012	Prospective registry	133/3290	IS	IS; or Cardiovascular events	1 year
Suzuki et al	2012	Prospective multicentre	NA/3324	2 week ≤ IS ≤6 months	IS	NA
van Wijk et al (LiLAC)	2005	Prospective RCT with retrospective follow-up	NA/2362	TIA or minor stroke	Long-term vascular events	10 years

Abbreviations: events – n; Sample size – N; Not Applicable – NA; IS – Ischaemic Stroke; TIA – Transient Ischaemic Stroke

3.3.1 Qualitative assessment of development studies

3.3.1.1 Cohort characteristics

Studies which collect data prospectively have a lower risk of information and selection biases for both baseline data and outcome events occurring during follow-up. Most studies used prospectively collected data, though 4/12 did not (Ay et al., 2010, Kernan et al., 1991, van Wijk et al., 2005, Kamouchi et al., 2012) one of which (van Wijk et al., 2005) used prospective trial data but included retrospective events obtained beyond the trial's original follow-up period. Few, 4/12, development studies recruited patients consecutively from routine practice (Ay et al., 2010, Pezzini et al., 2009, Putaala et al., 2010, Kamouchi et al., 2012). Loss of patients to follow-up often occurs when studies last for long time periods. Most, 9/12, (Ay et al., 2010, Dhamoon et al., 2007, Kernan et al., 1991, Putaala et al., 2010, Sumi et al., 2012, Suzuki et al., 2012, van Wijk et al., 2005, Stahrenberg et al., 2013, Kamouchi et al., 2012) development studies reported loss to follow up; and where it could be calculated, 7/8, (Dhamoon et al., 2007, Kernan et al., 1991, Putaala et al., 2010, Sumi et al., 2012, Suzuki et al., 2012, van Wijk et al., 2005, Weimar et al., 2009, Stahrenberg et al., 2013, Kamouchi et al., 2012) rates of loss were small (less than 5%).

The most frequent variables included in multivariable clinical prediction models were: age, history of TIA or stroke, history of hypertension, and diabetes (Appendix B Table 3-5). Five articles (Dhamoon et al., 2007, Kernan et al., 1991, Pezzini et al., 2009, Stahrenberg et al., 2013, Kamouchi et al., 2012) defined all predictors, three (Ay et al., 2010, Putaala et al., 2010, Sumi et al., 2012) defined only some, and four (Diener et al., 2005, Kernan et al., 2000, Suzuki et al., 2012, van Wijk et al., 2005) did not define any. Most articles defined outcome adequately, though three failed to define the outcome and/or the duration of follow-up (Diener et al., 2005, Kernan et al., 2000, Suzuki et al., 2012).

Figure 3-3 Aspects of model development

Missing baseline data occur frequently when collecting information from patients. A complete case analysis using only those patients with complete baseline data risks selection bias and loss of information. Five of the development studies (Kernan et al., 1991, Kernan et al., 2000, Pezzini et al., 2009, Suzuki et al., 2012, van Wijk et al., 2005) reported missing data; four (Kernan et al., 1991, Pezzini et al., 2009, Suzuki et al., 2012, van Wijk et al., 2005) of which stated the impact a complete case analysis had on the derivation sample size. No attempts were made to impute missing data.

3.3.1.2 Statistical methods

Most investigators collect more potential predictors than will be included in their final model. Data dependent methods (e.g., univariate selection or stepwise selection) are often used to select a few '*important variables*' from those available to develop a prediction model. This can lead to over-fitted models that perform over-optimistically in their development datasets which may be impossible to replicate in external evaluation (Steyerberg, 2009). Most of the studies used data dependent variable selection methods: stepwise selection (2/12) (Kernan et al., 1991, Putaala et al., 2010); univariate significance tests (4/12) (Ay et al., 2010, Dhamoon et al., 2007, Pezzini et al., 2009, Kamouchi et al., 2012); and further reduction of univariate selection by inspection of multivariable significance (2/12) (Suzuki et al., 2012, van Wijk et al., 2005). Three modifications of pre-existing prediction models were identified with new predictors chosen by clinical justification (Kernan et al., 2000, Sumi et al., 2012, Stahrenberg et al., 2013). One study gave no description of how variables were selected (Diener et al., 2005).

Models were derived using Cox proportional hazard regression (9/12) (Ay et al., 2010, Dhamoon et al., 2007, Kernan et al., 1991, Kernan et al., 2000, Pezzini et al., 2009, Putaala et al., 2010, Suzuki et al., 2012, van Wijk et al., 2005, Stahrenberg et al., 2013) or multivariable binary logistic regression (3/12) (Diener et al., 2005, Sumi et al., 2012, Kamouchi et al., 2012). Most studies presented their models as point scores (7/12) by rounding regression coefficients (Ay et al., 2010, Diener et al., 2005, Kernan et al., 1991, Kernan et al., 2000, Sumi et al., 2012, Pezzini et al., 2009, Kamouchi et al., 2012). The categorisation of a continuous predictor results in the loss of information. The majority of studies categorised continuous predictors (8/12) (Ay et al., 2010, Dhamoon et al., 2007, Diener et al., 2005, Kernan et al., 1991, Kernan et al., 2000, Pezzini et al., 2009, Sumi et al., 2012, Kamouchi et al., 2012), only one of which gave some clinical justification for the cut-points chosen (Sumi et al., 2012). The remaining four studies used a mixture of categorised and continuous variables (Putaala et al., 2010, Suzuki et al., 2012, van Wijk et al., 2005, Stahrenberg et al., 2013).

Internal evaluation methods use the model development data to provide optimism-corrected estimates of model performance. Few authors internally assessed their models performance (3/12) using cross-validation methods (Ay et al., 2010, Kamouchi et al., 2012, Sumi et al., 2012).

A common rule of thumb used in prediction model literature is the 'ten events per tested variable' (10 EPV) rule. The median total sample size across the twelve development studies was 1132 (IQR 522 to 3123). Where reported (9/12) the median number of events was 73 (IQR 60 to 102). Only one of the five studies where the EPV could be calculated had more than the minimum recommended EPV of 10 events per parameter (Sumi et al., 2012).

3.3.2 Meta-analysis of evaluation studies

3.3.2.1 Study characteristics

The ESSEN Stroke Risk Score (ESRS) (Diener et al., 2005), the Stroke Prognosis Instrument II (SPI-II) (Kernan et al., 2000), the Recurrence Risk Estimator at 90 days (RRE-90) (Ay et al., 2010) and the Life Long After Cerebral ischemia (LiLAC) (van Wijk et al., 2005) were externally evaluated in eleven different studies. Four additional evaluations were identified amongst the model development studies (Ay et al., 2010, Kernan et al., 2000, Sumi et al., 2012, Stahrenberg et al., 2013) giving fifteen evaluation cohorts: five evaluations of the ESRS (Alvarez-Sabin et al., 2008, Fitzek et al., 2011, Sumi et al., 2012, Weimar et al., 2009, Weimar et al., 2008); three of the SPI-II (Kernan et al., 2000, Navi et al., 2011, Wijnhoud et al., 2010); five head-to-head comparisons of the ESRS and the SPI-II (Ay et al., 2010, Chandratheva et al., 2011, Meng et al., 2011, Weimar et al., 2010, Weimar et al., 2012, Stahrenberg et al., 2013); one head-to-head comparison of the ESRS and the RRE-90 (Maier et al., 2013); and one comparing the ESRS, the SPI-II and the LiLAC models (Weimar et al., 2010) (Table 3-2).

The median sample size in the 15 evaluation cohorts was 1286 (IQR 619 to 5004). Various combinations of events and follow-up periods were used yielding 49 specific AUROCC values for extraction (Appendix BC Table 3-6). Where the effective sample size could be determined the median size was 86 (IQR 58 to 134).

Table 3-2 Characteristics of 15 evaluation studies assessing the performance of prediction models for recurrent vascular events after ischaemic stroke

Study	Model(s)	Study design	Baseline event	Outcome(s)	Country	Follow-up
Alvarez-Sabin (2008)	ESRS	Prospective multicentre	IS	IS/new vascular events	Spain	6 months
Ay (2010)	ESRS/SPI-II	Single centre	IS	IS	America	90 days
Chandratheva (2011)	ESRS/ SPI-II/ ABCD2	Prospective multicentre	TIA or Stroke	IS	UK	90 days
Fittzek (2011)	ESRS	Prospective single centre	TIA or acute IS	IS	Germany	13.4 (5.9) months[1]
Kernan (2000)	SPI-II	Prospective RCT	TIA or IS	IS/death	NA	2 years
Maier (2013)	ESRS/ABCD2/ RRE-90	Prospective single centre	IS	IS/CV death	Germany	8 (6 to 11) days[2]
Meng (2011)	ESRS/SPI-II	Multicentre registry	TIA or IS	IS & any vascular event	China	1 year
Navi (2011)	SPI-II	Community based	IS	IS and/or death	California	1 year
Stahrenberg (2013)	ESRS/SPI-II	Prospective observational	IS	CV events/ total mortality	Germany	1 year
Sumi (2012)	ESRS	Prospective registry	IS	IS/MI/CV death	Japan	1 year
Weimar (2010)	ESRS/SPI-II/ LiLAC	Prospective multicentre	TIA or non-disabling IS	IS/stroke & CV death	Germany	1 year
Weimar (2008)	ESRS	Prospective multicentre	TIA or acute IS	IS/stroke & CV death	Germany	17.5 (0.88) months[1]
Weimar (2009)	ESRS	Prospective registry	TIA or IS	IS, MI, and CV death	NA	1 year
Weimar (2012)	ESRS/SPI-II	Prospective multicentre	IS	IS, MI, and CV death	Germany	1 year
Wijnhound (2010)	SPI-II	Prospective single centre	TIA or minor IS	IS, MI & vascular death / fatal or nonfatal IS	Netherlands	2 years

Abbreviations: Ischaemic stroke – IS; Cardiovascular – CV. [Note: 1 – mean (SD); 2 – median (IQR)]

3.3.2.2 Statistical performance

A meta-analysis of the AUROCC values observed in evaluation studies of the ESRS and the SPI-II models are presented in Figure 3-4. The median sample size in eleven pooled studies for the ESRS was 1727 (IQR 700 to 3292) and for the SPI-II was 1372 (IQR 572 to 6486). Where the effective sample size could be determined the median size was 78 (IQR 52 to 124) and 96 (IQR 51 to 964) respectively.

The pooled AUROCC values for the ESRS was 0.60 (95% CI 0.59 to 0.62) and for the SPI-II was 0.62 (95% CI 0.60 to 0.64) (Figure 3-4). Six head to head comparisons of the ESRS and the SPI-II were identified. Four of these (Stahrenberg et al., 2013, Meng et al., 2011, Weimar et al., 2010, Weimar et al., 2012) (the other two (Ay et al., 2010, Chandratheva et al., 2011) used much shorter follow-up periods) were pooled to calculate the AUROCC estimates: 0.61 (95% CI 0.58 to 0.64) with 95% PI (0.29 to 0.93) and 0.62 (95% CI 0.59 to 0.66) with 95% PI (0.23 to 0.99) respectively for the ESRS and the SPI-II scores. One evaluation study for the RRE-90 score estimated an AUROCC of 0.72 (95% CI 0.64 to 0.80) (Maier et al., 2013) and another of the LiLAC score estimated an AUROCC of 0.65 (95% CI 0.61 to 0.70) (Weimar et al., 2010). Two evaluations of the ABCD2 score (Johnston et al., 2007) were identified (Chandratheva et al., 2011, Maier et al., 2013). Although the ABCD2 score did not meet the inclusion criteria of this systematic review these evaluations suggests generalisability to a broader range of patients suffering ischaemic strokes for the prediction of recurrent stroke (Figure 3-4). Only one study assessed the calibration of the SPI-II score which found it to be good but only after re-calibration (Wijnhoud et al., 2010). There was no evidence for small study (i.e., publication) bias (see Appendix B Figure 3-5).

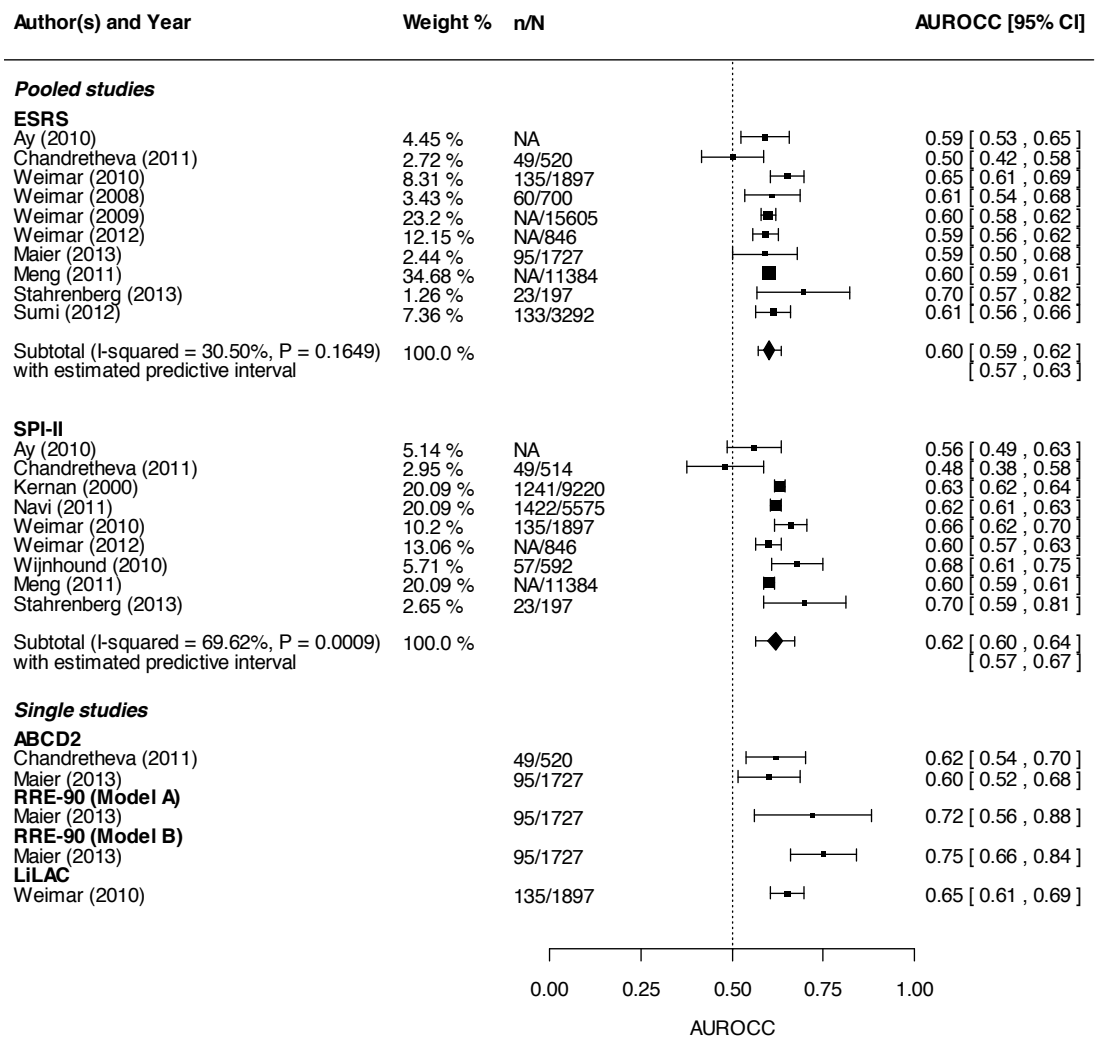


Figure 3-4 Meta-analysis of the AUROCC values for the ESRS and the SPI-II (Note: percentage weights are from random effects analysis)

3.3.2.3 Sensitivity analyses

Some evaluation studies deviated from the original derivation outcome definitions and follow-up times. However, often multiple AUROCC values were provided (see Appendix B Table 3-6). A series of sensitivity analyses for the ESRS and the SPI-II are provided in the table below (Table 3-3) meta-analysing those AUROCC values reported for specific outcome definitions. Small differences in the AUROCC values were identified, but note that after the first row in this table the AUROCC values are no longer comparable between the two models since they relate to different endpoints.

Table 3-3 Sensitivity analyses for meta-analysis of AUROCCs for ESRS and SPI-II

Sensitivity analyses	No. studies	AUROCC	95% CI	95% PI
All studies				
ESRS	10	0.60	0.59 to 0.62	0.57 to 0.63
SPI-II	9	0.62	0.60 to 0.64	0.56 to 0.67
Outcome as per development				
ESRS	8	0.59	0.57 to 0.61	0.28 to 0.91
SPI-II	7	0.62	0.61 to 0.64	0.39 to 0.86
Follow-up time as per development				
ESRS	5	0.61	0.59 to 0.62	0.43 to 0.78
SPI-II	1	0.68	0.61 to 0.75	NA
Follow-up and outcome as per development				
ESRS	4	0.60	0.57 to 0.63	0.24 to 0.95
SPI-II	1	0.68	0.61 to 0.75	NA

3.4 Discussion

There were four externally evaluated clinical prediction models for the prediction of recurrent stroke and myocardial infarction after stroke (the ESRS, the SPI-II, the RRE-90 and LiLAC) each having modest but similar discrimination (AUROCC 0.60 to 0.72).

Design weaknesses may explain the modest performance observed in external evaluation: the categorisation of continuous variables may have caused a loss in predictive power; data-dependent variable selection may have led to models over-fit to the data; and the cohorts were generally too small for reliable model development, in fact only one study achieved more than the recommended minimum of 10 EPV. The cohorts used to develop the models had weaknesses that are frequent in epidemiological studies: there were missing baseline data; whether the recruited patients were representative of those seen in routine clinical practice was uncertain; some data were collected retrospectively; and most cohorts did not record all potentially predictive variables. Despite the differences in the derivation of the ESRS, the SPI-II and the LiLAC, they all discriminated similarly (and modestly) between patients with and without recurrent stroke (Weimar et al., 2010). The ESRS and the SPI-II have four predictors in common (age, history of TIA or stroke, diabetes, and blood pressure). Three head to head comparisons demonstrated a relative difference in AUROCC which did not exceed 2% (Meng et al., 2011, Weimar et al., 2010, Weimar et al., 2012).

It is not known whether the discrimination of any prediction model for recurrent stroke is better or worse than a clinicians' informal prediction, therefore their utility in clinical practice remains unclear. Despite direct comparisons between formal and informal predictions perhaps providing the most robust argument for or against the use of statistical prediction model, they remain rare. For instance, there are many prediction rules for poor outcome or disability after stroke (Veerbeek et al., 2011) but few have been tested against clinicians' informal predictions (Counsell et al., 2004). In Chapter 4 a comparison will be made between formal statistical prediction and

informal clinicians' predictions of recurrent vascular events in an observational cohort.

3.4.1 Implications for research

This review identified a number of areas which could, if adopted, potentially improve the discrimination of future clinical prediction models for recurrent stroke or myocardial infarction: (i) using all the available information from a cohort by avoiding the categorisation of continuous predictors, and using multiple imputation of missing data where a complete case analysis would exclude a significant proportion of the cohort; (ii) reporting regression coefficients (i.e., prior to any transformation) to allow more accurate evaluation of models in independent cohorts. Point score models are probably obsolete as more precise predictions can easily be obtained using applications accessed via mobile computers at the bedside. There are too many proposed models in clinical practice to remember them all, and it is only sensible that they should be available electronically; and finally, (iii) measuring whether newly identified predictors, (e.g. blood markers or imaging techniques) add to the accurate classification of patients over more easily measured variables, for example using the net reclassification improvement (NRI) (Pencina et al., 2008, Stahrenberg et al., 2013).

A number of methodological decisions in model development may lead to clinical prediction models that make less accurate predictions (Bouwmeester et al., 2012). It is for this reason that an agreed set of guidelines in model development and reporting in healthcare would be helpful to developers and users of clinical prediction models alike (Collins, 2011).

The impact of a clinical prediction model on clinical practice should ultimately be assessed through a randomised trial where the use of a prediction model forms part of the intervention. A well designed impact study should incorporate an assessment of patient outcome as well as cost-effectiveness (Moons et al., 2012a). This systematic review suggests moderate discrimination of recurrent stroke and MI when applying the ESRS or the SPI-II. Either of these models could be selected, re-

calibrated for a given population, and assessed using an impact study thus quantifying the overall worth of the model.

3.4.2 Limitations of the study

Assessing the quality of studies of predictive models is difficult, and there is no widely agreed set of guidelines. This is likely to become an increasing problem as such studies are frequent, and very likely will begin to influence practice.

No scoring tool was used to rank the development studies. However, it is debatable whether such a tool should be used in assessing model quality. For instance it may be the case that despite methodological flaws in development the model performs well in practice. Indeed, given the quality observed through this systematic review, by assigning poor scores to quality of model development has the potential to cause an excess of 'non-starters'. In addition it is not clear how each item should be weighted: intuitively a score of '1' should have the same impact across the full range of quality items, however, it is not clear that this would strictly be the case. Finally, model performance assessed through external evaluation in new data is by far the more robust way to rate the quality of the model.

The electronic search was overly sensitive and returned a small number of relevant articles. Additional searches of the 'grey' literature were not performed, although this would not likely have resulted in many more articles. This is an unfortunate artefact of poor indexing, as no Medical Subject Heading (MESH) terms for clinical prediction model papers currently exist. Extensive use of forward citation searching in Google Scholar was made in an attempt to work around these limitations.

3.5 Appendix A: electronic search strategy

Table 3-4 Electronic search term implemented in Medline and EMBASE

Medline

1. cerebrovascular disorders/ or basal ganglia cerebrovascular disease/ or exp brain ischemia/ or carotid artery diseases/ or carotid artery thrombosis/ or carotid stenosis/ or cerebrovascular accident/ or exp brain infarction/ or exp hypoxia-ischemia, brain/ or exp intracranial arterial diseases/ or exp "intracranial embolism and thrombosis"/
2. ((brain or cerebr$ or cerebell$ or vertebrobasil$ or hemispher$ or intracran$ or intracerebral or infratentorial or supratentorial or middle cerebr$ or mca$ or anterior circulation) adj5 (isch?emi$ or infarct$ or thrombo$ or emboli$ or occlus$ or hypoxi$)).tw.
3. (isch?emi$ adj6 (stroke$ or apoplex$ or cerebral vasc$ or cerebrovasc$ or cva or attack$)).tw.
4. 1 or 2 or 3
5. ((risk or predictive or prediction or statistical or cox or logistic or survival or multivariate or multivariable or hazard$) and (prediction or model$ or equation or rule or calculator)).tw.
6. cox proportional hazard model$.ab. or cox proportional hazard model$.ti. or cox proportional-hazard$.ab. or cox proportional-hazard$.ti.
7. 5 or 6
8. 7 and 4
9. Child/ or ethnic groups/ or *Depression/ or economics/ or *caregivers/
10. 8 not 9
11. (Bibliography or Editorial or Letter or News).pt.
12. 10 not 11
13. limit 12 to 1980-current
14. limit 13 to human.

EMBASE

1. cerebral artery disease/ or cerebrovascular accident/ or stroke/ or vertebrobasilar insufficiency/ or wallenberg syndrome/ or exp brain infarction/ or exp brain ischemia/ or exp occlusive cerebrovascular disease/ or cerbrovascular disease/ or exp carotid artery diseases/

2. ((brain or cerebr$ or cerebell$ or vertebrobasil$ or hemispher$ or intracran$ or intracerebral or infratentorial or supratentorial or middle cerebr$ or mca$ or anterior circulation) adj5 (isch?emi$ or infarct$ or thrombo$ or emboli$ or occlus$ or hypoxi$)).tw.
3. (isch?emi$ adj6 (stroke$ or apoplex$ or cerebral vasc$ or cerebrovasc$ or cva or attack$)).tw.
4. 1 or 2 or 3
5. ((risk or predictive or prediction or statistical or cox or logistic or survival or multivariate or multivariable or hazard$) and (prediction or model$ or equation or rule or calculator)).tw.
6. cox proportional hazard model$.ab. or cox proportional hazard model$.ti. or cox proportional-hazard$.ab. or cox proportional-hazard$.ti.
7. 5 or 6
8. 7 and 4
9. Child/ or ethnic group/ or depression/ or health economics/ or caregiver/
10. 8 not 9
11. (Bibliography or Editorial or Letter or News).pt.
12. 10 not 11
13. limit 12 to 1980-current
14. limit 13 to human

3.6 Appendix B: Extra tables and figures

Table 3-5 illustrates the variables used in each of the 12 model development studies. Some presented multiple models per publication making for a total of 28 models which could be extracted.

Figure 3-5 shows graphical assessment of publication bias using contour-enhanced funnel plots.

Table 3-6 lists the various additional AUROCC values that could be extracted from the model evaluation studies.

Table 3-5 Overview of predictors considered in the 12 development papers for recurrent vascular events after ischaemic stroke

Predictors	Total per study	Ay (2010)		Dhamoon (2007)		Diener (2005)	Kamouchi (2012)		Kernan (1991)	Kernan (2000)	Pezzini (2009)			Putaala (2009)				Stahrenberg (2013)			Sumi (2012)		Suzuki (2012)	van Wijk (2005)					
Model per publication		**1**	**2**	**1**	**2**	**1**	**1**	**2**	**1**	**1**	**1**	**2**	**3**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**1**	**2**	**1**	**1**	**2**	**3**	**4**	**5**	**6**
Demographics																													
Age	**9**			•	•	•	•	•	•	•				•	•	•	•				•	•	•	•	•	•	•	•	•
Gender	**4**													•	•	•	•				•	•	•	•	•	•	•	•	•
Social Factors																													
Smoking	**4**					•	•	•							•	•	•				•	•							
Past medical history																													
TIA/stroke	**7**	•	•			•	•	•		•					•		•				•	•	•						
Stroke in family	**1**										•		•																
CAD	**2**			•	•					•																			
PAD	**3**					•									•	•	•				•	•							
CHF/Heart failure	**3**			•	•					•				•	•		•												
CHD	**3**								•						•	•	•												
MI	**5**				•	•									•		•				•	•		•	•	•			
Diabetes	**8**					•	•	•	•	•				•	•	•					•	•	•	•	•	•	•	•	•
Chronic kidney disease	**1**						•	•																					

Table 3-5 Continued from previous page

Predictors	Total per study	Ay (2010)	Dhamoon (2007)	Diener (2005)	Kamouchi (2012)	Kernan (1991)	Kernan (2000)	Pezzini (2009)	Putaala (2009)	Stahrenberg (2013)	Sumi (2012)	Suzuki (2012)	van Wijk (2005)					
Past medical history cont.																		
Other cardiovascular disease (not MI/AF)	**3**			•	•	•						•	•					
Peripheral vascular surgery	**1**												•	•				
Intermittent claudication	**1**												•	•	•	•	•	•
Details of stroke																		
Stroke not TIA	**3**					•	•							•	•		•	•
Lacunar subtype	**1**		•															
Nonlacunar subtype	**1**				•	•												
Embolic subtype	**1**		•															
Subtype (not SAO)	**1**										•	•						
Admission CCS	**2**	•	•						•	•	•	•						
Migraine with aura	**1**							•		•								
Amaurosis Fugax	**1**													•				
Risk score: ESRS	**1**									•								
Risk score: SPI-II	**1**									•								
NIHSS	**1**									•								

Table 3-5 Continued from previous page

Predictors	Total per study	Ay (2010)	Dhamoon (2007)	Diener (2005)	Kamouchi (2012)	Kernan (1991)	Kernan (2000)	Pezzini (2009)	Putaala (2009)	Stahrenberg (2013)	Sumi (2012)	Suzuki (2012)	van Wijk (2005)					
General examination cont.																		
Waist circumference/ Obesity	**3**								•		•	•	•					
Hypertension or BP	**9**			•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Hyperlipidemia	**1**											•						
AF	**3**		•	•	•				•									
mRS	**1**											•						
Paresis	**1**																•	•
Dysarthia	**1**													•	•		•	•
Dyslipidemia	**1**								•									
Vertigo	**1**													•	•			
CT/MRI/ECG																		
Multiple infarcts of different ages	**1**	•																
Simultaneous infarcts in different circulations	**1**	•																
White matter lesions	**1**														•			•
Any infarct	**1**														•			•

Table 3-5 Continued from previous page

Predictors	Total per study	Ay (2010)	Dhamoon (2007)	Diener (2005)	Kamouchi (2012)	Kernan (1991)	Kernan (2000)	Pezzini (2009)	Putaala (2009)	Stahrenberg (2013)	Sumi (2012)	Suzuki (2012)	van Wijk (2005)	
CT/MRI/ECG cont.														
Q wave on ECG	**1**												•	
Negative T wave	**1**												•	
ST-depression	**1**												•	•
Genetic factors/biomarkers														
FV_{G1691A}	**1**							•	•					
TT677 MTHFR	**1**							•	•					
$PT_{G20210A}$	**1**							•	•					
hsTropT	**1**									• • •				

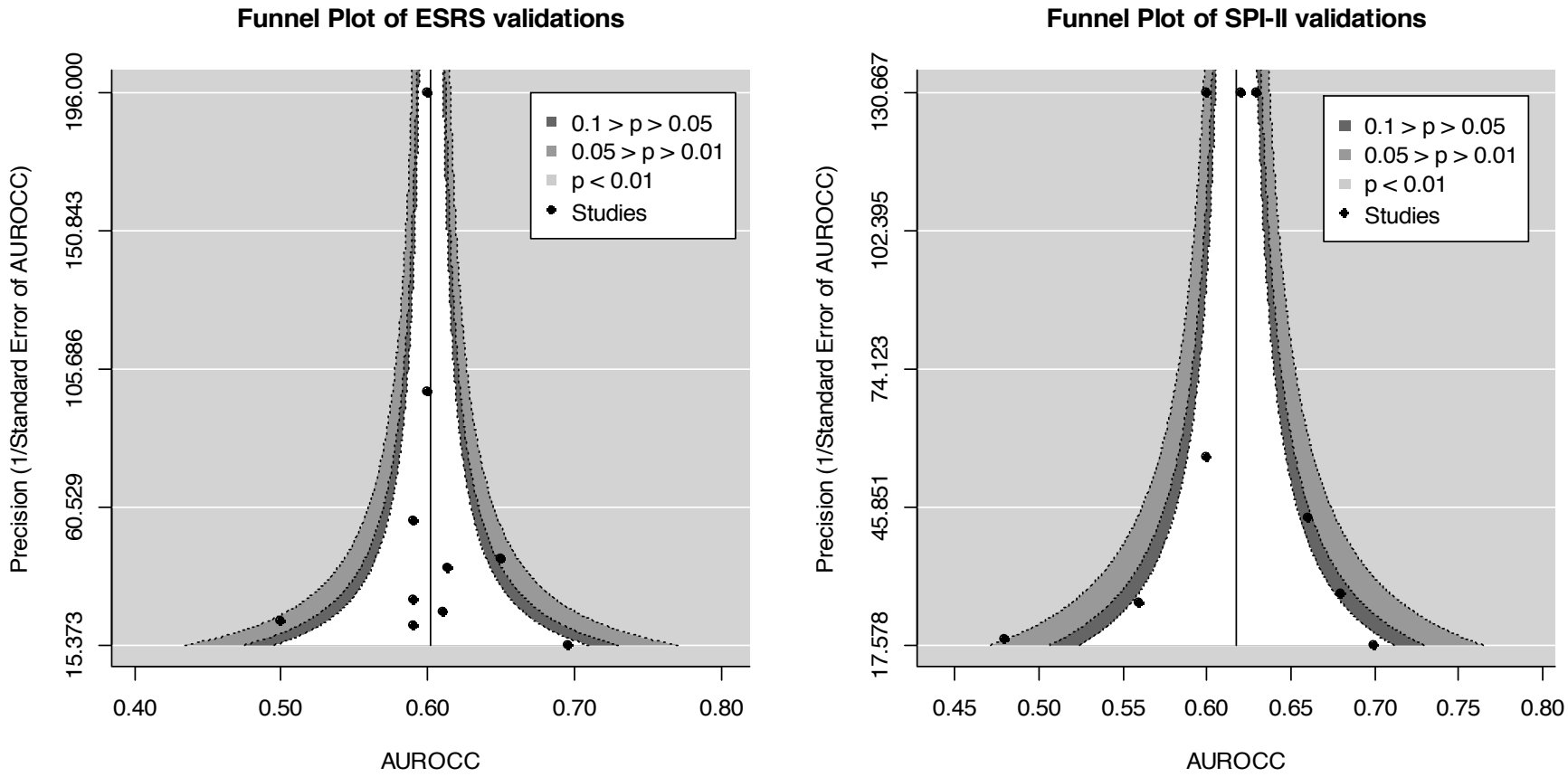


Figure 3-5 Contour-enhanced funnel plots for assessing publication bias in ESRS and SPI-II evaluation studies

Table 3-6 Discrimination metrics for externally evaluated models (n = number of events, N = total sample size, CV = cardiovascular, IS = ischaemic stroke)

Study	AUROCC (95%CI)	n/N	Outcome	Additional AUROCC with 95% CI
ESRS				
Ay (2010)	0.59 (0.53 to 0.66)	NA	IS	-
Chandretheva (2011)	0.50 (0.42 to 0.59)	49/520	IS	0.49 (0.35 to 0.62), for 7 day events
Fitzek (2011)	0.59 NA	76/723	IS	-
Weimar (2010)	0.65 (0.60 to 0.69)	135/1897	IS or CV death	0.62 (0.57 to 0.67) for IS
Weimar (2008)	0.61 (0.54 to 0.69)	60/700	IS or CV death	0.56 95%CI. NA for IS
Weimar (2009)	0.60 (0.58 to 0.62)	NA/15605	Nonfatal IS/MI or CV death	0.56 (0.53 to 0.58) for IS
Weimar (2012)	0.59 (0.56 to 0.63)	NA/846	Nonfatal IS/MI or CV death	0.62 (0.59 to 0.65) fatal/non-fatal IS
Maier (2013)	0.59 (0.50 to 0.68)	95/1727	IS	0.50 (0.42 to 0.58), for 7 day events
Meng (2011)	0.60 (0.59 to 0.61)	NA/11384	Nonfatal IS/MI or CV death	0.59 (0.58 to 0.60) for IS, for baseline IS only: 0.60 (0.59 to 0.62) for nonfatal IS/MI or CV death and 0.60 (0.57 to 0.61) for IS
Stahrenberg (2013)	0.695 (0.567 to 0.822)	23/197	IS or CV death	0.744 (0.575 to 0.912), for all-cause mortality
Sumi (2012)	0.613 (0.564 to 0.661)	133/3292	IS/MI or CV death	0.604 (0.554 to 0.654) for IS
SPI-II				
Ay (2010)	0.56 (0.49 to 0.64)	NA	IS	-
Chandretheva (2011)	0.48 (0.39 to 0.60)	49/514	IS	0.50 (0.37 to 0.64), for 7 day events
Kernan (2000)	0.63 (0.62 to 0.65)	1241/9220	IS or death	-
Navi (2011)	0.62 (0.61 to 0.64)	1422/5575	IS or death	0.55 (0.51 to 0.59) stroke; 0.64 (0.62 to 0.66) death

Table 3-6 Continued from previous page

Study	AUROCC (95%CI)	n/N	Outcome	Additional AUROCC with 95% CI
SPI-II continued				
Weimar (2010)	0.66 (0.61 to 0.70)	135/1897	IS or CV death	0.65 (0.60 to 0.70) for IS
Weimar (2012)	0.60 (0.57 to 0.64)	NA/846	Nonfatal IS/MI or CV death	0.56 (0.53 to 0.60) fatal/non-fatal IS
Wijnhound (2010)	0.68 (0.61 to 0.75)	57/592	IS/MI or CV death	0.64 (0.56 to 0.72) fatal/non-fatal IS
Meng (2011)	0.60 (0.58 to 0.61)	NA/11384	Nonfatal IS/MI or CV death	0.59 (0.58 to 0.61) for IS, for baseline IS only: 0.61 (0.59 to 0.62) for nonfatal IS/MI or CV death and 0.60 (0.58 to 0.62) for IS
Stahrenberg (2013)	0.699 (0.587 to 0.810)	23/197	IS or CV death	0.708 (0.549 to 0.867), for all-cause mortality
ABCD2				
Chandretheva (2011)	0.62 (0.54 to 0.70)	49/520	IS	0.64 (0.53 to 0.74), for 7 day events
Maier (2013)	0.60 (0.52 to 0.69)	95/1727	IS	0.60 (0.53 to 0.67), for 7 day events
LiLAC				
Weimar (2010)	0.65 (0.61 to 0.70)	135/1897	IS or CV death	0.64 (0.59 to 0.69) for stroke
RRE-90 A/B				
Maier (2013)	0.72 (0.56 to 0.88)	95/1727	IS	0.58 (0.46 to 0.70), for 7 day events
Maier (2013)	0.75 (0.66 to 0.84)	95/1727	IS	0.68 (0.58 to 0.77), for 7 day events

Chapter 4: Predictions of recurrent stroke and MI after ischaemic stroke

Background and summary

This chapter presents an assessment of models for the prediction of recurrent stroke or MI after stroke identified via a systematic review of the available literature between 1980 till 19th of April 2013 in a new prospective cohort of stroke patients. A unique record of clinicians' informal predictions raised the opportunity to compare formal and informal predictions. It is concluded that neither method out performs the other and that discrimination of events from non-events is poor.

4.1 Introduction

The occurrence of a recurrent stroke or an MI during a patient's recovery following an initial ischaemic stroke could have important implications for both the state of independent living the patient can expect to return to as well as their chance of long term survival. If it were possible to reliably determine the absolute risk of a recurrent vascular event then the treating clinicians could use this to implement a suitable preventative strategy or else direct certain admissions to specialised stroke centres. To make the best use of medical resources it is therefore as important to identify those at low risk as it is to identify those at high risk: such evidence could inform important aspects across recovery (Lemmens et al., 2013). There are two ways that recurrence free survival may be predicted: (i) using clinicians' predictions, i.e., a gut feeling based on clinical experience; or (ii) using some clinical prediction model, combining a number of prognostic risk factors whose importance is suitably weighted within a single equation. Each method will draw from a similar pool of available information, but may use it in different ways. In the previous chapter it was seen that a number of clinical prediction models already exist. The conclusion was that by and large model development was poor and that many of the weaknesses

common amongst available models developed for other diseases persist (Bouwmeester et al., 2012). Whilst model development is important, provided that the model performs well in practice, it is of little concern how well the derivation was undertaken (Collins and Le Manach, 2013). This is not to devalue the importance of using robust and methodologically sound model development techniques as in the absence of external evaluation; a close inspection of model development is by far the best way of judging which model to back. The performance of a clinical prediction model is best assessed though in a new prospective cohort of patients by investigators with no vested interest in how the results turn out (Collins et al., 2014). This chapter has two aims: (i) to test the performance of existing clinical prediction models in a new cohort of first time stroke patients; and (ii) to assess whether there are any important differences between those predictions made by clinicians and those made by statistical models.

4.2 Materials and Methods

The previous chapter covered many important aspects of model development asking general qualitative questions of all those models identified. In what follows a more detailed description is provided for each of the models whose performance could be evaluated in the data made available for analysis in this chapter. A summary of the available data will be given following which a discussion as to how it compares to the derivation data from which each model came.

4.2.1 Predicting recurrent vascular events

Five of the twelve identified clinical prediction models could be assessed in the Edinburgh Stroke Study. All were derived as Cox Proportional Hazard models (PHMs) except for the ESRS which was derived using binary logistic regression. Most (3 out of 5) were presented as a point score.

4.2.1.1 The ESSEN Stroke Risk Score (ESRS)

The ESSEN Stroke Risk Score (ESRS) was developed using prospectively collected trial data from the Clopidogrel versus Aspirin in Patients at Risk of Ischemic Events (CAPRIE) trial (Diener et al., 2005). Interestingly, whilst this model has had numerous external evaluations its development has been poorly reported.

4.2.1.2 The second Stroke Prognosis Instrument (SPI-II)

The second Stroke Prognosis Instrument (SPI-II) was developed by Kernan *et al*. and was offered as an improvement on the original SPI-I, which, the authors suggest, may have been based on too small a sample (142 patients with 38 events) (Kernan et al., 1991, Kernan et al., 2000). The authors explored the inclusion of new predictors based on existing literature which Kernan *et al*. tested in the Women's Estrogen for Stroke Trial (WEST). These new variables were fit as a multivariable Cox PHMs where only those variables that achieved a P-value less than or equal to 0.1 were included. The estimated regression coefficients were rounded to produce a point score. The authors quote identifiable risk groups (i.e., low, medium and high) and provided the Kaplan-Meier estimates of stroke or death within each of these strata.

4.2.1.3 The 90 day Recurrence Risk Estimator (RRE-90)

Ay *et al*. developed the 90 day Recurrence Risk Estimator (RRE-90) to predict ischaemic stroke patient's risk of recurrent stroke by 90 days with a recurrence defined as a visible area of new infarction through imaging (Ay et al., 2010). The patients included in the study were obtained retrospectively between 2003 and 2006 from a single center in the US. They developed their model in a sample of 1458 patients of which 60 suffered a recurrent stroke. Two models were published: Model A, a *clinical-based model*; and Model B, *Clinical-and imaging-based model*. A process of univariable significance was used to establish which of the important predictors should make it into a Cox PHM with the obtained regression coefficients rounded to produce a simple point score. Model A included history of TIA or stroke and patient's Admission CCS (Causative Classification of Stroke System) subtype. Model B included: multiple infarcts of different ages; simultaneous infarcts in different circulations; multiple acute infarcts; isolated cortical infarcts; history of TIA or stroke; and admission CCS subtype. The authors provided the Kaplan-Meier estimated risk from the derivation data split into five distinct score strata. In this chapter only Model A will be evaluated due to an unavailability of patient baseline data in the ESS.

4.2.1.4 Putaala's model

Putaala *et al*. investigated the risk of recurrent vascular events in a young cohort of 807 stroke patients aged from 15 to 49 (Putaala et al., 2010). They cited the need for a better understanding of the etiological differences in the risk of recurrence within a younger population of stroke patients as the primary motivation for introducing a new model. They developed four separate prediction models using two distinct outcomes: nonfatal/fatal ischaemic stroke, and a composite outcome of any arterial event. They incorporated covariates using two strategies: individual clinical risk factors and stroke etiology; or a simple count of clinical risk factors, as well as age, gender and stroke etiology. A backward stepwise selection process was used to choose the most important risk predictors.

4.2.1.5 Dhamoon's model

Dhamoon *et al*. developed a model for myocardial infarction or vascular death after a first time ischaemic stroke using a prospective population cohort (Dhamoon et al., 2007). They studied a sample of 655 patients of which 102 suffered the event of interest by 5 years. Cox PHMs were used to explore univariable and multivariable associations amongst historical and physiological risk factors as well as patients' stroke etiological subtype. Those covariates identified as *significant* at the univariable level with a P-value <0.1 made it into a multivariable fit at which point further selection was undertaken using a stepwise algorithm. The authors published two models: one included age, CAD, CHF and AF; and another included age, MI, CAD without MI, CHF, AF and stroke subtype. In this chapter only the first of these will be evaluated due to an unavailability of patient baseline data in the ESS.

4.2.1.6 Clinicians' predictions in the Edinburgh Stroke Study

Informal predictions of recurrent events in the Edinburgh Stroke Study were made on patients seen at outpatient clinics only. These were made as an absolute percentage quantifying the perceived risk of the event. It was possible to obtain some information regarding thirteen of the clinicians making predictions for 542 (94%) of the patients. Eight were neurologists (62%) and five were stroke physicians (38%); seven were in training (54%) and six were fully trained (46%). The median number of patients seen per clinician was seven (ranging from 1 to 217).

Table 4-1 Overview of the five clinical prediction models

Model	ESRS (2005)	SPI-II (2000)	RRE-90 (2010)	Putaala (2010)	Dhamoon (2007)
Source population	Prospective RCT	Prospective RCT	Retrospective single center	Prospective single center	Prospective population cohort
Outcome	Recurrent stroke	Stroke or death	Recurrent stroke	Fatal or nonfatal ischaemic stroke	MI or vascular death
Time window	1 year	2 years	90 days	5 years	5 years
No. predictors	8	7	2	6	4
Predictors					
Demographics	Age	Age		Age Gender	Age
Clinical characteristics		Stroke not TIA			
Risk factors	Prior TIA	Prior stroke	Prior stroke or TIA	Prior TIA	History of Atrial Fibrillation
	Arterial hypertension	Severe hypertension		Heart failure	History of coronary heart failure
	Prior myocardial infarction	Coronary heart failure		Diabetes Mellitus	History of coronary artery disease
	Diabetes Mellitus	Diabetes Mellitus			
	Cerebrovascular disease, excluding MI	Coronary artery disease			
	Peripheral Artery Disease				
	Smoking				
Stroke etiology			Admission CCS subtype	TOAST classification	

Abbreviations: Trial of Org 10172 in Acute Stroke Treatment – TOAST; Causative Classification of Stroke System – CCS

4.2.2 The Edinburgh Stroke Study (ESS)

The Edinburgh Stroke Study (ESS) was a prospective observational study of stroke patients admitted to the Western General Hospital in Edinburgh between April 2002 and May 2005. Patients were followed for a minimum of one year. Details on the study's design have been published elsewhere (Jackson et al., 2009). In brief, clinicians were asked to use *gut-feeling* to estimate the absolute risk of a recurrent stroke or a vascular event – that is, stroke, MI or vascular death – within one year of the initial stroke amongst those seen as outpatients (N=671). The definition of recurrent stroke required a period of neurological stability of at least 24 hours between the initial stroke and the recurrent stroke whilst simultaneously excluding any other potential causes of neurological deterioration. Myocardial infarction in follow-up was defined either through autopsy or else through evidence of at least two of the following signs: symptoms of myocardial ischemia, e.g., chest pain; enzyme changes associated with MI, e.g., elevated troponin levels; and ECG changes suggesting new ischemia. Probable MI was defined as any abrupt deaths with no alternative explanation. There were few missing baseline data (see Table 4-2) therefore complete case data are used throughout this chapter.

4.2.3 Cohort Comparability

Important differences may exist between the development and evaluation populations. Baseline data were extracted from each of the five prediction models as far as was possible, including variables not used in the final proposed models. No baseline data could be extracted for the ESRS due to poor reporting of its development. Variables included in the final models are highlighted in Table 4-2 in bold. ESS did not classify stroke according to the Causative Classification of Stroke System (CCS); although a record of classification as per the Trial of Org 10172 in Acute Stroke Treatment (TOAST) algorithm could be manipulated to a form that closely resembled the CCS. Considerable differences are noted when comparing the prevalence of the observed patient characteristics in ESS to those patient characteristics obtained for each of the five prediction model cohorts (see Table 4-2). The development population used by Putaala *et al*. was younger by design and it may therefore be anticipated that far fewer would have a poor medical history.

Table 4-2 Baseline characteristics of the five prediction model cohorts and the evaluation cohort. Note '*' denotes the mean, else the median is presented

Variable	ESS cohort	Missing	ESRS	SPI-II	RRE-90 (Model A)	Putaala	Dhamoon
Total patients in study	1257	-	-	525	1458	807	655
Outpatients	53% (671)	-	-	-	-	-	-
Outcomes							
Recurrent stroke (<90 days)	4% (52)	-	-	-	60	-	-
Recurrent stroke (within one year)	8% (102)	-	-	90	-	-	-
Myocardial infarction (within one year)	2% (28)	-	-	-	-	-	102
Any vascular event (within one year)	22% (274)	-	-	-	-	-	-
Baseline characteristics							
Age (median, *mean*)	74, *72**		-	-	*72 (no event), 74 (event)*	***42****	***70****
< 65 years	26% (321)	-	-	**27% (141)**	-	-	-
65 – 75 years	46% (577)	-	-	-	-	-	-
> 75 years	29% (359)	-	-	-	-	-	-
> 70 years	61% (770)	-	-	**57% (301)**	-	-	-
Male	51% (644)	-	-	0% (0)	54% (783)	**63% (508)**	45% (292)
History of hypertension	54% (683)	1	-	**71% (373)**	-	39% (314)	70% (454)
History of diabetes mellitus	13% (158)	-	-	**31% (165)**	24% (349)	7% (54)	45% (295)
Previous Myocardial Infarction	28% (350)	-	-	24% (124)	-	4% (31)	16% (106)
Heart Failure	6% (80)	3	-	6% (29)	-	5% (29)	14% (90)
Current or prior atrial fibrillation	22% (271)	2	-	16% (83)	30% (431)	4% (30)	11% (72)
Peripheral Artery Disease	8% (96)	5	-	-	-	2% (17)	22% (141)
Current or Ex-Smoker < 12 months	66% (820)	22	-	-	21% (310)	47% (377)	54% (349)
Prior TIA or ischaemic stroke	31% (391)	3	-	34% (178)	**8% (110)**	-	-
Admission CCS subtype							
Large Artery Atherosclerosis	8% (104)	-	-	-	**23% (338)**	**8% (68)**	-
Cardio-aortic embolism	14% (171)	-	-	-	**25% (358)**	**19% (151)**	-
Small Artery occlusion	21% (258)	-	-	-	**11% (153)**	**14% (112)**	-
Other Causes	8% (105)	-	-	-	**6% (83)**	**26% (209)**	-
Undetermined Causes	49% (619)	-	-	-	**36% (526)**	**33% (267)**	-

4.3 Outcomes in follow-up for ESS

After a follow-up period of one year 8% of patients (102/1257) had suffered a *recurrent stroke* whilst 22% (274/1257) had suffered the broader *any vascular* event. Note that each of the identified prediction models was developed to predict patient outcomes with somewhat different definitions (Table 4-1). However, the previous chapter illustrated that it is commonplace for investigators to test the generalisability of a given prediction model. Indeed from a practical perspective it is of interest to assess whether prediction models may be generalised beyond those patients seen in the initial development cohort. For this reason, each of the clinical prediction models described in Table 4-1 will be evaluated in the ESS using the following outcome definitions: (i) recurrent stroke within one year; (ii) any vascular event within one year; and, as far as is possible (iii) the original development outcome.

4.3.1 Discrimination: formal versus informal

Clinicians' informal prediction of recurrent stroke and any vascular events was available for those seen as outpatients (N=671) in the ESS. All comparisons between informal and formal prediction are therefore made within this subset. The median time from stroke onset to assessment for outpatients was 19 days (IQR 12 to 28). Formal tests using the method described by DeLong *et al*. are provided comparing the ROC curves obtained from each clinical prediction model to that of the clinicians (Figure 4-1 and Table 4-3) under the null hypothesis of no true difference in AUROCC (DeLong et al., 1988). Contrasting informal clinicians' predictions (clinical gestalt) with formal clinical prediction models suggests that when discriminating between those who do suffer an event (either recurrent stroke or any vascular event) from those who do not, both methods of prediction perform poorly yet similarly (all DeLong tests P-value>0.05). When testing these clinical prediction models in all patients (Figure 4-2 and Table 4-3), each of the models demonstrated an improvement in AUROCC discriminating *any vascular event* compared to the *recurrent stroke* outcome. Using the outcome definitions as originally defined in model derivation generally resulted in the highest discrimination, except for the ESRS which performed best when discriminating between those with and those without any vascular events across one year follow-up.

Table 4-3 Discriminative performance of informal clinicians' predictions (clinical gestalt) and clinical prediction models in the Edinburgh Stroke Study

	Recurrent stroke				Any vascular event				Original development outcome		
Prediction Method	**n/N**	**AUROCC**	**95% CI**	**P-value***	**n/N**	**AUROCC**	**95% CI**	**P-value***	**n/N**	**AUROCC**	**95% CI**
ESS outpatients only											
Clinical gestalt	40/575	0.53	0.44 to 0.63	-	63/574	0.56	0.48 to 0.64	-	-	-	-
ESRS (2005)	50/664	0.56	0.48 to 0.64	0.8409	80/664	0.57	0.50 to 0.63	0.8751	-	-	-
SPI-II (2000)	50/669	0.58	0.49 to 0.66	0.9664	80/669	0.59	0.52 to 0.66	0.5385	-	-	-
RRE-90 (2010)	50/671	0.61	0.52 to 0.69	0.4404	80/671	0.59	0.53 to 0.66	0.7266	-	-	-
Dhamoon (2007)	50/668	0.60	0.52 to 0.68	0.5043	80/668	0.61	0.54 to 0.67	0.2005	-	-	-
Putaala (2010)	50/669	0.48	0.39 to 0.57	0.4995	80/669	0.56	0.49 to 0.63	0.9422	-	-	-
ESS outpatients and inpatients											
ESRS (2005)	101/1224	0.54	0.49 to 0.60	-	256/1224	0.62	0.59 to 0.66	-	101/1224	0.54	0.49 to 0.60
SPI-II (2000)	102/1253	0.53	0.48 to 0.59	-	274/1253	0.63	0.59 to 0.67	-	274/1253	0.63	0.59 to 0.67
RRE-90 (2010)	102/1254	0.58	0.52 to 0.64	-	273/1254	0.59	0.56 to 0.63	-	52/1254	0.59	0.51 to 0.67
Dhamoon (2007)	102/1253	0.57	0.51 to 0.63	-	274/1253	0.68	0.64 to 0.71	-	205/1253	0.73	0.69 to 0.76
Putaala (2010)	102/1247	0.50	0.44 to 0.55	-	269/1247	0.65	0.61 to 0.68	-	269/1247	0.65	0.61 to 0.68

* Obtained from the DeLong *et al*. test of two correlated ROC curves, carried out using the pROC command in R (Robin et al., 2011)

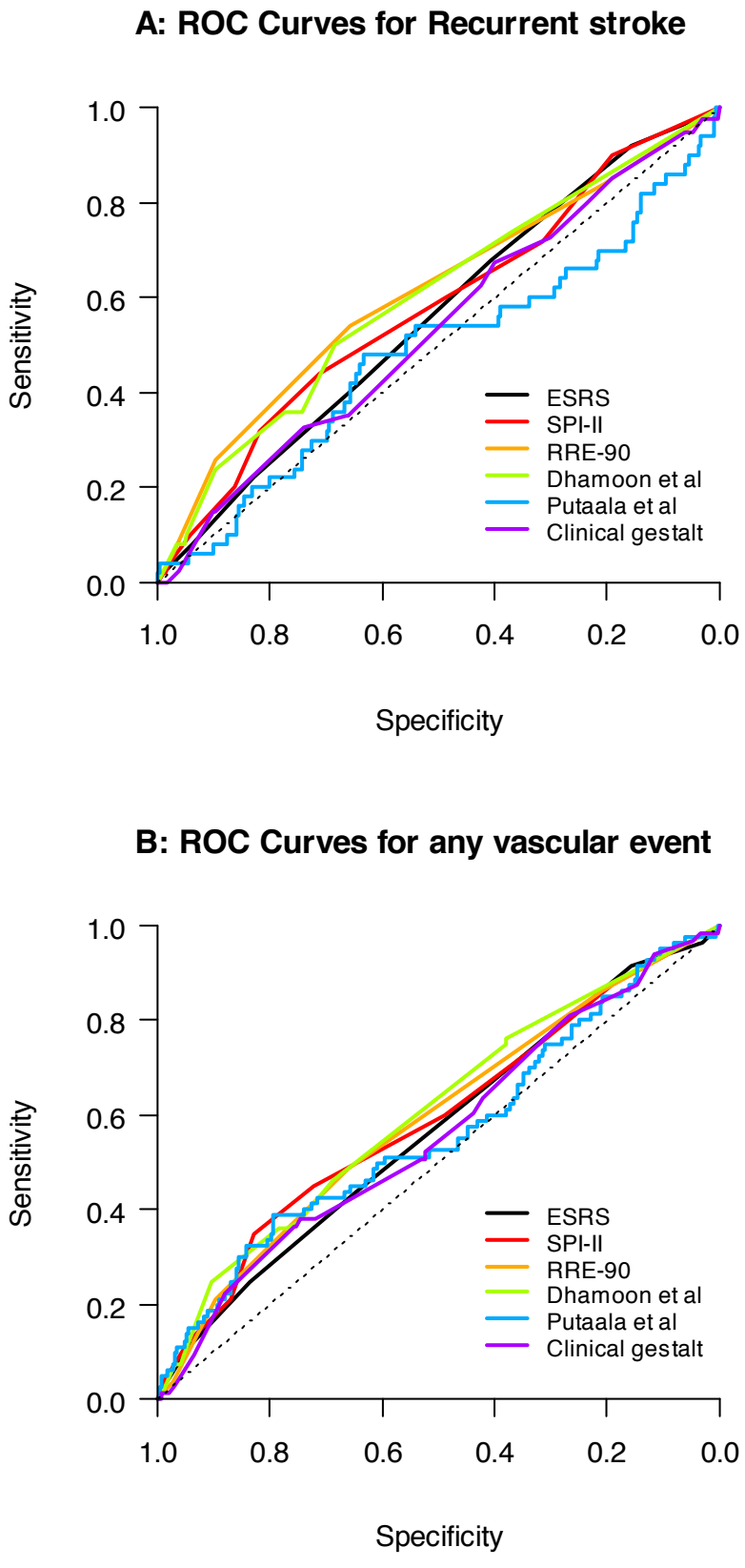


Figure 4-1 Receiver Operating Characteristic (ROC) curves contrasting discrimination as achieved by formal and informal methods: (A) recurrent stroke; and (B) any vascular event

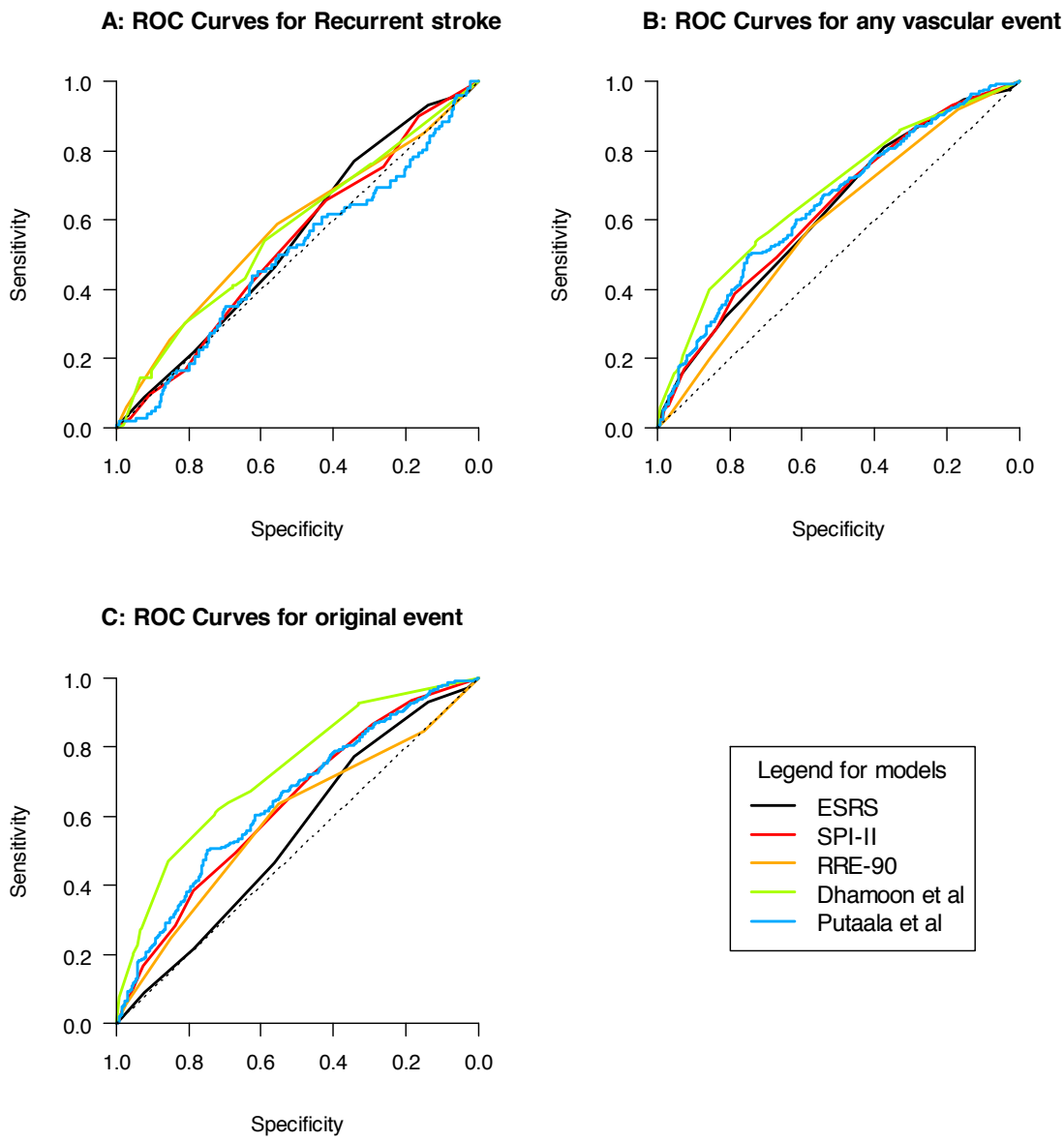


Figure 4-2 Receiver Operating Characteristic (ROC) curves contrasting discrimination as achieved by various formal clinical prediction models: (A) recurrent stroke; (B) any vascular event; and (C) original event as specified in development.

4.3.2 Discrimination: updated meta-analysis

A meta-analysis of AUROCC values obtained through the systematic review of model evaluation studies was presented in the previous chapter for two models: the Essen Stroke Risk Score (ESRS) and the Stroke Prognosis Instrument (SPI-II). Table 4-4 below summarises the updated pooled estimates obtained after a random effects meta-analysis when including the AUROCC estimates obtained from the Edinburgh Stroke Study. The AUROCC values pooled in Chapter 3 were those estimated for the discrimination of any vascular event, therefore the corresponding estimates in the ESS were used to update the previous estimates (Table 4-3). The prediction intervals based on ten and nine previous model evaluation studies respectively were accurate in capturing a future estimate of AUROCC for the ESRS and SPI-II.

No qualitative or indeed practical quantitative improvement in the pooled estimate was detected after updating the previous meta-analysis.

Table 4-4 Updated meta-analysis of ESRS and SPI-II discrimination

Analysis	ESRS	SPI-II
Original Meta-analysis		
Number of studies	10	9
AUROCC with 95% CI	0.60 (0.59 to 0.62)	0.62 (0.60 to 0.64)
95% PI	0.57 to 0.63	0.57 to 0.67
Q-statistic, P-value*	12.95, 0.1649	26.34, 0.0009
Updated with ESS data		
AUROCC with 95% CI	0.60 (0.59 to 0.62)	0.62 (0.60 to 0.64)
95% PI	0.57 to 0.63	0.57 to 0.67
Q-statistic, P-value*	14.42, 0.1548	26.88, 0.0015

* Obtained by comparing the *Q*-statistic to a χ^2 distribution with $k - 1$ *df*

4.3.3 Calibration: formal versus informal

Frequently the final model presented in publication is in the form of a point score making no or little reference to the absolute risk observed in the derivation data. In general, baseline risk is the constant term in any regression model with all predictor variables set equal to zero – this is the intercept in the case of a logistic regression or some estimate for the baseline hazard in the case of the Proportional Hazards model. A point score involves some rounding of the estimated regression coefficients, perhaps after being scaled by some constant. Risk categories may then be presented (e.g., low, medium and high), which was the case for the SPI-II and the RRE-90 models thus providing some information about the absolute risk associated with the original derivation data. This information was utilised to generate calibration plots (plotting predicted risk against observed risk in new data) and associated calibration slopes and calibration intercepts in the ESS – though the small number of category levels will have an impact on the accuracy with which the calibration slope and intercept can be estimated which are transformed on to the logit scale and fit as continuous (see Table 4-5 and Figure 4-3). Contrasting the calibration of informal and formal prediction in the ESS data, neither prediction method performs particularly well with considerable systematic biases and slopes less than one. Information about absolute predicted risk could not be obtained for the three remaining models and were therefore not assessed.

Table 4-5 Calibration of clinical gestalt and clinical prediction models in the ESS (note, all quoted intervals are 95% CIs)

Outcome	Clinical gestalt	SPI-II	RRE-90
Recurrent stroke			
n/N	40/575	50/669	50/671
Intercept	0.97 (0.64 to 1.29)	-1.13 (-1.43 to -0.84)	0.24 (-0.07 to 0.55)
Slope	0.18 (-0.39 to 0.76)	0.25 (-0.20 to 0.69)	0.39 (0.13 to 0.66)
Any vascular event			
n/N	63/574	80/669	80/671
Intercept	1.49 (1.23 to 1.76)	-0.59 (-0.83 to -0.35)	0.85 (0.59 to 1.10)
Slope	0.39 (-0.08 to 0.86)	0.37 (0.01 to 0.72)	0.32 (0.10 to 0.54)
Original development outcome			
n/N	-	274/1253	52/1254
Intercept	-	0.07 (-0.07 to 0.21)	-0.62 (-0.91 to -0.33)
Slope	-	0.62 (0.42 to 0.83)	0.26 (0.00 to 0.52)

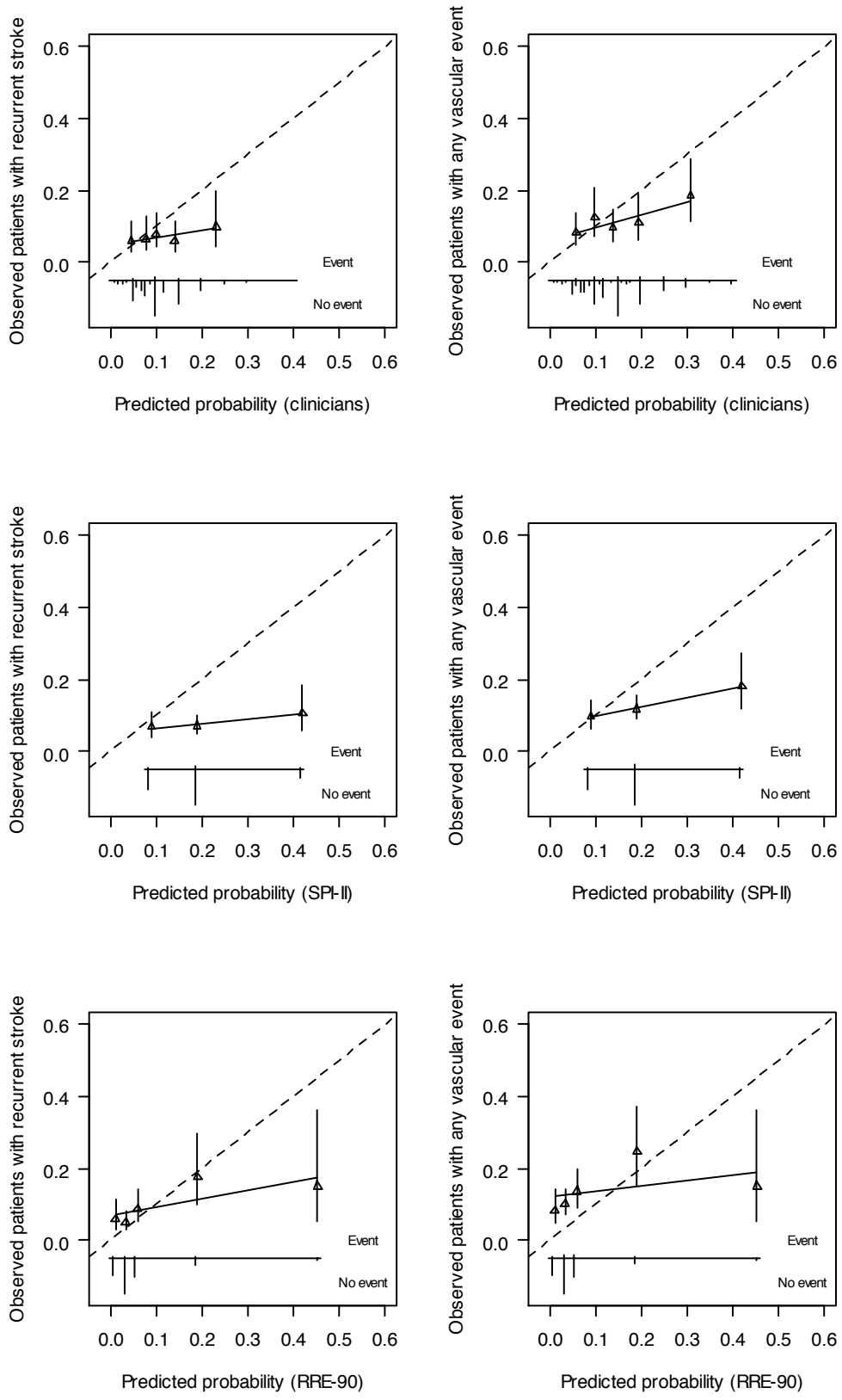


Figure 4-3 Calibration plots for informal and formal prediction of recurrent stroke and any vascular event. Note: 95% CIs are calculated using the Wilson method (Wilson, 1927).

4.4 Discussion

It is well recognised that for a clinical prediction model to successfully enter into clinical practice it must have a clear and intended application ultimately resulting in better patient care and improve the chance of patient recovery (Moons et al., 2012a, Altman et al., 2009, König et al., 2007). Evidence of good performance (i.e., high discrimination and accurate calibration) must be demonstrated in a number of external cohorts to convince clinicians (the intended consumers of the research) that the model has the potential to be useful in prognostication. Model performance in the ESS cohort was similar to that seen in previous model evaluation studies (see Table 4-4 and Chapter 3). There appears to be a sufficient body of evidence to conclude that the ESRS and the SPI-II achieve modest discrimination between those with and those without vascular events in follow up amongst those cohorts already assessed.

One aspect that typically receives less attention is that of comparing predictions made by statistical models to the clear alternative, *expert predictions from experienced clinicians*. The comparison of prediction models for vascular events to informal clinicians' predictions presented in this chapter appears to be one of the few studies in this area. Of those patients observed as part of the Edinburgh Stroke Study only those seen as outpatients had clinicians' predictions. Prediction models performed similarly to clinicians' predictions in this subset but the separation of events from non-events using the same models appeared to be better when applied to inpatients and outpatients combined (Table 4-3) or alternatively when used on inpatients only (see Table 4-6 below). The median time from stroke onset to assessment for inpatients was 2 days (IQR 1 to 4). It must be understood though that the patient seen at an outpatient clinic differs from the patient admitted to hospital. The incidence of recurrent stroke in the ESS for instance was 9% and 7% for inpatients and outpatients respectively; similarly, the incidence of any vascular event was 33% and 12%, suggesting differing case-mix. It was not possible from these data to say how well clinicians performed at discriminating between events for those admitted to the inpatients wards.

Table 4-6 Discriminative performance of clinical prediction models in the 586 inpatients. Note all intervals are 95% CIs

	Recurrent stroke		Any vascular event		Original development outcome	
Model	**n/N**	**AUROCC**	**n/N**	**AUROCC**	**n/N**	**AUROCC**
ESRS (2005)	51/560	0.51 (0.44 to 0.59)	176/560	0.62 (0.57 to 0.67)	51/560	0.51 (0.44 to 0.59)
SPI-II (2000)	52/584	0.47 (0.40 to 0.54)	194/584	0.61 (0.57 to 0.66)	194/584	0.61 (0.57 to 0.66)
RRE-90 (2010)	52/583	0.53 (0.45 to 0.62)	193/583	0.53 (0.49 to 0.58)	32/583	0.56 (0.46 to 0.66)
Dhamoon (2007)	52/585	0.53 (0.45 to 0.62)	194/585	0.67 (0.62 to 0.72)	167/585	0.71 (0.66 to 0.75)
Putaala (2010)	52/578	0.53 (0.45 to 0.61)	189/578	0.64 (0.60 to 0.69)	189/578	0.64 (0.60 to 0.69)

It is important to place these findings in a broader context. Historically it has been shown that predictions made by statistical models are more accurate than those made by experts irrespective of the medical speciality (Grove et al., 2000). With regards to prediction in stroke other researchers have compared formal and informal prediction methods. For example, there are numerous prediction rules for poor outcome or disability after stroke, however Counsell *et al*. demonstrated that the sensitivity and specificity achieved by clinical prediction models was comparable to that achieved by clinicians (Veerbeek et al., 2011, Counsell et al., 2004). Studies of other clinical conditions have demonstrated better discrimination achieved via statistical prediction contrast with predictions made by an experienced clinician. For example, Dehing-Oberije *et al*. showed that the occurrence of radiation-induced dysphasia in patients with lung cancer could be better identified by a statistical model than clinicians' predictions (Dehing-Oberije et al., 2010).

The role of a clinical prediction model is not to replace the treating clinician but rather to act as an extra tool in the clinician's tool kit: it can help inform patient prognosis by offering an objective estimate of the absolute risk (Moons et al., 2012b, Kattan and Gerds, 2011). It is important to acknowledge that a clinical prediction model operates on a finite number of variates. There are many plausible scenarios under which nonsense predictions can be generated if applied *ignorant* to an important piece of extra information – information not formally incorporated within the statistical model. This has become known as Paul E. Meehl's *broken leg scenario* which should serve as an essential reminder that there is no replacement for intuition and interpretation (Salzinger, 2005). Provided that the obtained predictions are suitable at the point of application though, accurate clinical prediction models could play an important role within a range of clinical settings, for example: automating certain clinical processes; educating more junior doctors; as well as generating enriched patient recruitment to a clinical trial.

Finally, before applying a model in practice it is important to investigate whether the model fits the proposed setting. It should be recognised that calibration is a joint property of both model and data. A model may be said to be *well calibrated* for certain settings but it would be ill-advised to assume that this is therefore a general

property of that model (Vickers and Cronin, 2010). Indeed, predictions of the same outcome from different models can often have poor agreement due to differing case mix and baseline risk, in which case appropriate updating must be implemented (Steyerberg et al., 2005, Steyerberg, 2009). Calibration of the SPI-II and the RRE-90 were partly explored in the ESS, although a limited availability of development information means that these data should be interpreted with some caution.

Chapter 5: Predictions of functional outcome after ischaemic stroke

Background and summary

Prediction of functional outcome can be achieved formally through clinical prediction models or informally using expert clinical experience. This chapter compares each method for the prediction of functional outcome in a single observational cohort of stroke patients. It is found that clinical predictions models are at least as good as informal doctor's predictions at discriminating between good and bad functional outcome after ischaemic stroke.

5.1 Introduction

Recent ischaemic stroke patients, their families and their doctors would like an accurate prediction of patient recovery (i.e., a poor or good functional outcome). Such a prediction is typically made informally at the bedside by a doctor using his or her clinical experience and expertise. Not every patient will want to know the same amount of prognostic information (Back and Arnold, 2006). Nevertheless, predictions (formal or informal) are required and will influence how the patient is treated.

It is not clear whether formal statistical models do better or worse than doctors at predicting outcomes for stroke patients. In Chapter 4 the prediction of recurrent stroke and vascular events was explored comparing formal and informal methods. Both were fairly poor at discriminating outcomes though neither method outperformed the other. A similar evaluation is presented in the current chapter with the focus on the prediction of death or disability.

It has been previously shown that when contrast with algorithms or objective statistical models experts generally perform poorly at prediction (Kahneman, 2011,

Grove et al., 2000). The first proper review of this topic came from the American psychologist Paul E. Meehl whose seminal text "Clinical versus statistical prediction", published now over half a century ago, highlighted that in many unique settings predictions made by clinical psychologists were frequently outperformed by quite simple statistical models (Meehl, 1954). Meehl speculated that the reason for such disparity stemmed from a tendency for experts to overcomplicate scenarios: something which a simple objective statistical formula avoids.

The accuracy of predictions made by experts in the clinic depends upon the condition under consideration. For example, a study of doctors predicting the survival of terminally ill patients systematically gave overly optimistic survival times (Christakis and Lamont, 2000). A study comparing the subjective predictions of nurses and doctors of mortality in patients seen at an intensive therapy unit discriminated between those dead and alive better than the Acute Physiology and Chronic Health Evaluation II (APACHE II) – a score of disease severity of those seen in intensive care units (Marks et al., 1991). In the context of stroke, Counsell *et al*. demonstrated that similar levels of sensitivity and specificity were found when contrasting predictions of functional outcomes made by clinicians and by statistical models (Counsell et al., 2004). An important dimension to take into consideration is the environment in which the predictions are made and whether this can be considered as *regular* and *consistent*. Simply put: *do repeat scenarios arise from which the experts can learn*?

Expert intuition is strongly related to the breadth of experiences and identifiable scenarios likely to be encountered. As analogy, consider the experienced chess player who benefits from the familiarity of moves learned through thousands of previous games. In the same way the expert will base his or her prediction on his or her previous experience. Kahneman and Klein dubbed this the *validity of the environment* which, if considered high, will reflect the extent to which learning through experience can benefit prediction (Kahneman and Klein, 2009).

The aims of this chapter are to: (i) determine whether or not the predictions of functional outcome made informally by doctors in a single hospital are better or

worse than the predictions made by statistical models; and (ii) to explore what attributes (patient or clinician) make an informal prediction more or less accurate.

5.2 Materials

5.2.1 The Edinburgh Stroke Study (ESS)

The Edinburgh Stroke Study (ESS) was introduced previously in Chapter 4. In brief, it was a prospective observational study of recent stroke or TIA patients recruited consecutively from the outpatient, inpatient and emergency departments of the Western General Hospital, Edinburgh, UK between April 2002 and May 2005 (www.dcn.ed.ac.uk/ess/protocol) (Jackson et al., 2008). The Edinburgh Stroke Study had four main objectives: (i) to investigate how risk factors differ across the subtypes of ischaemic stroke (Jackson et al., 2010); (ii) to explore how recurrent events differ across ischaemic stroke subtypes (Jackson et al., 2009); (iii) to assess what role *novel* genetic risk factors or biomarkers have in the process of prediction (Whiteley et al., 2009); and (iv) to assess the current methods of predicting death or dependency and recurrent vascular events.

5.2.1.1 Defining types of doctor

Doctors with varying levels of experience in stroke medicine were asked to predict the six month Oxford Handicap Scale (OHS) for their patients at presentation using their clinical experience. Clinical experience was classified broadly into seniority (i.e., fully trained in neurology or stroke medicine versus in training) and speciality (i.e., geriatrics/internal medicine versus neurology). This was achieved retrospectively with reference to the initials recorded on each clinical report form for those patients with a doctors' prediction. Colleagues of the doctors at the Western General Hospital in Edinburgh were able to identify who each of the initials belonged to.

5.2.1.2 Outcome definition

Functional outcome was measured using the Oxford Handicap Scale (OHS) six months from the initial stroke onset. The OHS is an ordinal outcome summarising patient disability and death across seven unique levels. Levels 1 through to 5 denote disability increasing in levels of severity, i.e., a greater reliance on others for

undertaking everyday tasks. A score of 6 is reserved for mortality and a score of 0 for *fully recovered*. In this chapter *poor functional outcome* at six months was defined as an OHS of 3 or more. All analyses were restricted to those patients with definite or probable ischaemic stroke. This was defined as a focal deficit of cerebral origin lasting for at least 24 hours, where brain imaging showed either positive evidence of cerebral infarction, or was normal or equivocal and the clinical syndrome was most in keeping with stroke. This definition excluded patients with haemorrhagic strokes because such patients likely differ in recovery to those with ischaemic stroke. Functional outcome was measured independently (i.e., not by the original clinician seen on admission) with the OHS at six months using a validated postal questionnaire. Non-responders were sent a repeat questionnaire. Patients were *flagged* for death with the General Register Office for Scotland, which provided information on the date, place and cause of death.

5.2.1.3 Baseline variable definitions

Doctors recorded baseline clinical variables using a standardised pro-forma. This included basic patient demographic variables (e.g., patient age, smoking status etc.) as well as more novel blood biomarkers (Whiteley et al., 2009). For the purpose of this chapter only those simple characteristics considered by the pre-existing models (Table 5-1) will be explored.

5.2.2 Pre-existing models

Five multivariable binary logistic regression models were identified (Table 5-1) (Appelros et al., 2003, Counsell et al., 2002, Lee et al., 2009, Reid et al., 2010, Weimar et al., 2004) from a previously published systematic review of models for the risk of poor functional outcome after stroke (Veerbeek et al., 2011). The linear predictor was calculated for each of these models using the associated regression coefficients reported by each author and in one case used the natural logarithm of the odds ratios to two decimal places (Appelros et al., 2003). Some models predicted *poor functional outcome* whilst others predicted *good functional outcome*. Since the latter is the inverse of the former, all models could be used to predict good or poor outcomes.

5.2.2.1 Counsell *et al.*

The Six Simple Variables (SSV) model (Counsell et al., 2002) was developed on 530 stroke patients seen within 30 days of stroke onset as part of the Oxfordshire Community Stroke Project (OCSP) between 1981 and 1986 (Bamford et al., 1990). Patients were defined as *independent* at six months if they achieved an OHS of less than three. Predictor selection was pragmatic: the authors included risk factors based on clinical importance and availability in clinical practice. The model included six easily obtained patient characteristics which summarised independence prior to the stroke and the impact of the stroke itself on patient ability at baseline. In the same publication the authors evaluated the performance of their model using an external dataset. Discrimination was good (AUROCCs ≥ 0.80) irrespective of: (i) time from stroke onset; (ii) type of stroke and (iii) history of previous stroke. There was a difference though when restricting to those patients seen as outpatients (0.65) compared to those seen as inpatients (0.83). Subsequent model evaluations have been undertaken suggesting good discrimination with the SSV model (AUROCCs ≥ 0.75) (Teale et al., 2012). Additionally, the SSV appears to have a degree of geographical generalisability with strong discrimination (AUROCC ≥ 0.90) amongst stroke patients in an observational study from China (Li et al., 2012).

5.2.2.2 Reid *et al.*

A model for independence (OHS<3) at six months (Reid et al., 2010) was developed using 538 stroke patients seen as part of the Stroke Outcomes Study (SOS) in Canada between 2001 and 2002. The purpose was to produce a model that outperformed the SSV by including CT scan information. Additivity and linearity assumptions were assessed. The final model fit used a stroke severity score which was derived from the EC/IC bypass study (The EC/IC Bypass Study group, 1985). This is rarely used in practice, and for the purpose of this chapter the NIHSS score was used in its place. The implications of this are discussed later.

5.2.2.3 Appelros *et al.*

Apperlos *et al*. aimed to investigate the risk factors associated with death, disability and recurrent stroke each within one year of first time stroke patients followed as part of a community based cohort study in Sweden (Appelros et al., 2003). They

assessed 377 stroke patients who had suffered a first time stroke within the period 1999 to 2000 with one year of follow-up. Various characteristics (i.e., demographic, medical history and stroke severity) were recorded for each patient at baseline. Patient age and stroke severity (NIHSS) were kept continuous. The authors used univariate tests to screen 13 risk factors excluding any variables not significant at the 5% level from entry into a multivariable fit. Binary logistic regression was used to model multivariable associations on the dichotomised outcome mRS≥3 at one year which included: patient age, stroke severity and history of heart failure. No intercept was reported precluding the assessment of calibration in external data sets. Additionally only the odds ratios could be obtained, although these could be transformed with the natural log to obtain the regression coefficients, there is an inevitable loss of precision.

5.2.2.4 Weimar *et al.*

Weimar *et al*. developed two prediction models for the separate prediction of incomplete recovery (defined as a Barthel index <95) or death after stroke and mortality (Weimar et al., 2004). They used 7238 patients with ischaemic stroke symptoms from a German stroke data bank between 1998 and 1999. This data source spanned a total of seven separate centres. They identified sixteen potentially important predictors by systematic review. This covered demographic aspects, patient medical history and the severity of the initial indexing stroke (e.g., NIHSS and other deficits). Any deviations from linearity for the continuous variables were assessed using fractional polynomials, although no improvement was made on the standard linear fits. They implemented both a forward and backward stepwise selection process excluding variables with a P-value > 0.001 and retaining those with a P-value ≤ 0.005. The authors evaluated the performance of their models in a separate dataset demonstrating that the model explained an important amount of variability in the data with an R^2 of 44%. Subsequent evaluation studies have demonstrated that this model achieves good discrimination of those with and those without the outcome of interest (AUROCC of 0.81) (König et al., 2008). The authors suggested that their models have potential use in the design of randomised trials and for adjustment in nonrandomised studies. In fact this model was used for this precise

purpose in the IST-3 trial to explore the effect of rtPA whilst adjusting for the predicted risk of the primary outcome based on the Weimar *et al*. model (The IST-3 collaborative group, 2012b).

5.2.2.5 Lee *et al.*

In their study of Taiwanese stroke patients, Lee *et al*. aimed to explore the underlying differences in risk profiles of those dead or dependent at six months (Lee et al., 2009). They recorded demographic, neurologic severity and lab based measurements in 533 first-ever ischaemic stroke patients recruited prospectively at a single hospital in Taiwan between October 2004 and November 2006. They developed a multivariable logistic regression using forward stepwise selection with the inclusion criterion P-value set to < 0.05.

5.2.3 Sensitivity Analyses

Clinical prediction models for risk of good functional outcome (OHS<3) perform differently when evaluated amongst patients seen in different settings (Counsell et al., 2002). Therefore all analyses were repeated within patients admitted to hospital and those seen at outpatient clinics. An alternative dichotomy of the Oxford Handicap Scale (OHS ≥ 2) was also explored as this often influences the results in stroke studies (Weisscher et al., 2008).

Table 5-1 Formal statistical prediction models for functional outcome.

Variables	Lee *et al.* (2009)	Appelros *et al.* (2003)	Weimar *et al.* (2004)	Counsell *et al.* (2002)	Reid *et al.* (2010)
Intercept			− 5.782	+ 12.340	+ 2.401
Age		+ 0.077	+ 0.049	− 0.051	− 0.049
Pre-stroke independence				− 2.744	+ 3.497
Living alone				+ 0.661	
Arm power				− 2.106	+ 1.402
Able to walk				− 1.311	
Normal GCS verbal				− 2.160	
NIHSS (stroke severity score)	+ 0.362	+ 0.285	+ 0.272		− 0.549
Heart failure		+ 1.099			
History of diabetes	− 2.296				
Total cholesterol	− 0.029				
Outcome	mRS>2 at six months	mRS≥3 at one year	BI<95 or dead	OHS≤2 at six months	OHS≤2 at six months
Source population	Taiwanese hospital cohort	Community based cohort of first ever strokes in Sweden	Stroke data bank of the German Stroke Foundation	OCSP community based incidence study	Consecutive patients enrolled in the Stroke Outcome Study
Additional comments	No intercept was provided therefore no calibration was possible	Coefficients estimated from the natural log of odds ratios reported to two decimal places		The SSV model scores 1 for presence and 2 for absence of risk factor	Stroke severity measured using a score adapted from the EC/IC bypass study

ABBREVIATIONS: modified Rankin Scale (mRS); the Oxford Handicap Scale (OHS); National Institutes of Health Stroke Scale (NIHSS); Glasgow Comma Scale (GCS); Barthel Index (BI); Six Simple Variables model (SSV); Extracranial/Intracranial (EC/IC) and the Oxford Community Stroke Project classification (OCSP)

5.3 Methods - measuring predictive accuracy

Functional outcome was recorded on a commonly used ordinal scale of death or disability, the Oxford Handicap Scale (OHS) at six months. Measuring the accuracy of formal or informal predictions can be approached in two ways: either (i) specify some dichotomous split; or else (ii) retain the ordinal structure. In this section each approach will be discussed, covering some methodological details before drawing some conclusion as to what impact this has on the interpretation in the context of the analyses.

5.3.1 Assessing accuracy on a dichotomous outcome

The informally predicted OHS and six month observed OHS scores were dichotomised into favourable (OHS<3) and unfavourable (OHS≥3) responses. A two by two cross classification of the form seen in Table 5-2 is produced. Two commonly encountered measures are the conditional probabilities: sensitivity and specificity which quantify the accuracy of a test (Zhou et al., 2008a). From Table 5-2 sensitivity describes the probability of correctly predicting an outcome given that the outcome occurs in follow up ($P\ (Pred = 1 \mid Obs = 1) = s_1 / n_1$) whereas specificity describes the probability of predicting absence of outcome given that the outcome is indeed absent in follow up ($P\ (Pred = 0 \mid Obs = 0) = r_0 / n_0$).

Table 5-2 Standard two by two cross-classification table

Predicted outcome (*Pred*)	Observed outcome (*Obs*)		Row total
	OHS<3 (*Obs* = 0)	OHS≥3 (*Obs* = 1)	
OHS<3 (*Pred*= 0)	r_0	r_1	m_0
OHS≥3 (*Pred* = 1)	s_0	s_1	m_1
Column total	n_0	n_1	N

For each binomial proportion, p, an associated 95% confidence interval can be produced. An asymptotic interval is commonly adopted which appeals to the normal approximation of the binomial distribution. However, as $p \in [0,1]$ such intervals can easily give illogical results with confidence intervals that exceed these bounds, especially when the estimated proportions are close to one or zero (Newcombe, 1998). In this chapter Zhou-Li confidence intervals (ZL CIs) have been used to provide confidence intervals for doctors' sensitivity and specificity (Zhou et al., 2008b). The ZL confidence interval has good coverage in circumstances where p is close to one or zero.

For each clinical prediction model (Table 5-1) the thresholds of predicted probability of poor functional outcome (OHS≥3) that had: (i) the same specificity; and (ii) the same sensitivity as the doctor's predictions were calculated. The model sensitivity was therefore calculated at the threshold of doctors' specificity and similarly the model specificity at the threshold of doctors' sensitivity. Informally, this equates to simply reading the sensitivity/specificity for fixed specificity/sensitivity off of the ROC curve as is illustrated in Figure 5-1. Each threshold was formally obtained using the R-package `pROC` which calculates the sensitivity and specificity obtained for all possible cut-points of the linear predictor (Robin et al., 2011).

This approach is used to address the question: given the accuracy of a doctor at successfully predicting OHS≥3 or ruling out OHS≥3 in favour of OHS<3 how does a clinical prediction model compare? A comparison between informal and formal predictions can then be made in a similar way as that seen in Counsell *et al.* (Counsell et al., 2004).

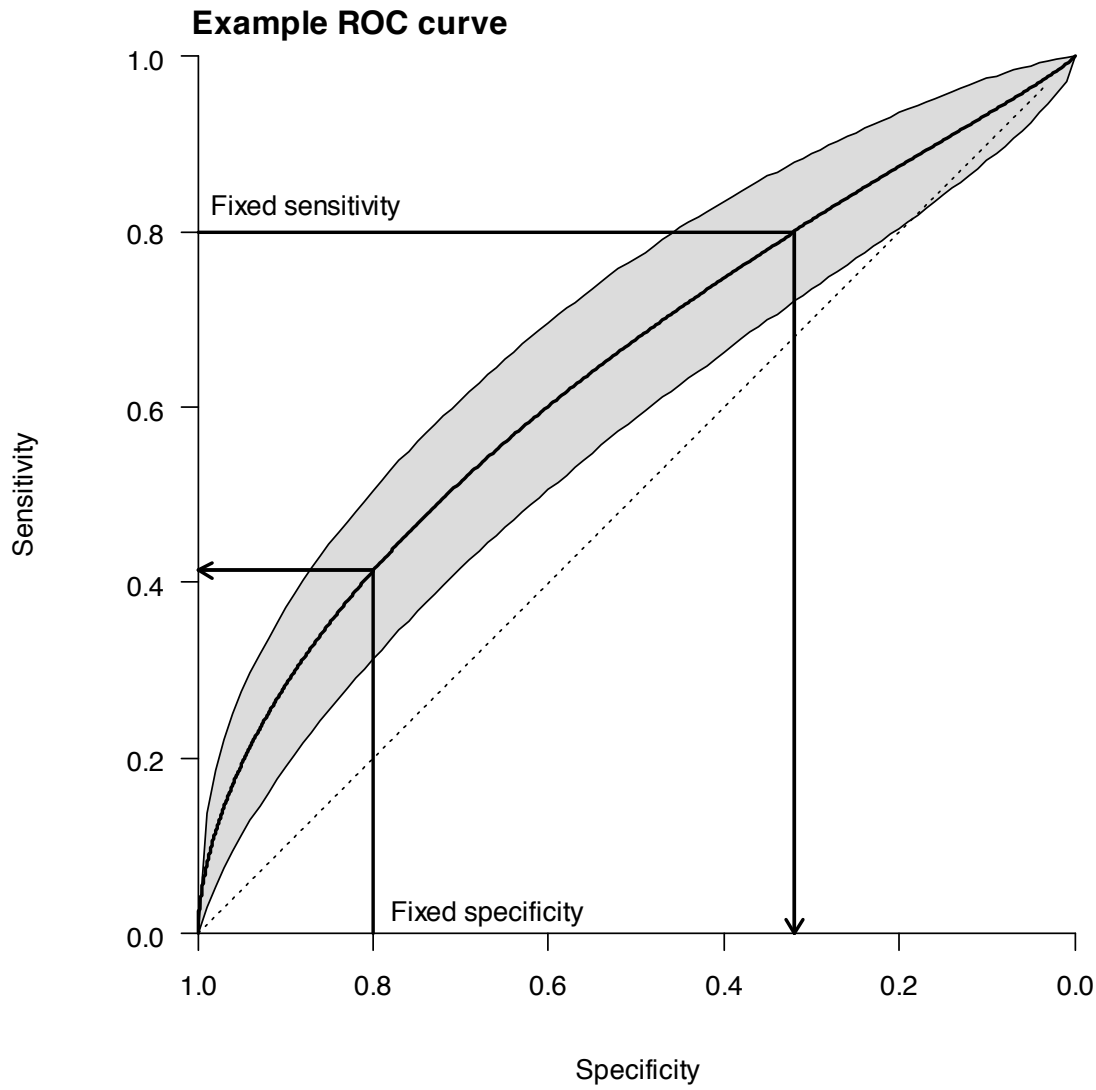


Figure 5-1 Example Receiver Operating Characteristic (ROC) curve with associated 95% confidence interval (grey shaded area)

5.3.2 Assessing accuracy on an ordinal outcome

5.3.2.1 Weighted kappa statistic

Cohen's kappa (κ) statistic is a commonly used measure of agreement between *raters* made on the same *subjects* (Graham and Jackson, 1993). It is calculated from the proportion of pairs which agree and the proportion of pairs that would be expected to agree by chance. By deducting the latter from the former the measure is said to be *corrected for chance agreement*.

$$\kappa = \frac{p_{obs} - p_{\exp}}{1 - p_{\exp}} \qquad \textbf{(5.1)}$$

A further specification is to adopt a weighting scheme awarding *partial credit* for near agreement (Cohen, 1968) the so called weighted kappa (κ_w).

Here p_{obs} and p_{exp} are the observed and expected weighted proportions respectively and are given by:

$$p_{obs} = \frac{1}{n_{\bullet\bullet}} \sum_i \sum_j w_{ij} n_{ij} \text{ and}$$

$$p_{\exp} = \left(\frac{1}{n_{\bullet\bullet}} \right)^2 \sum_i \sum_j (w_{ij} n_{i\bullet} n_{\bullet j}) \, .$$

Where, n_{ij} and w_{ij} denote the observed frequency and weight for the ijth cell respectively. Note that *bullet notation* has been adopted here to indicate the marginal sums of n_{ij} over the respective index (i.e., i or j) subsequently replaced by a *bullet point*. A broad range of different weighting schemes can be applied but the most commonly used are the squared weights,

$$w_{ij} = 1 - \frac{(i-j)^2}{(r-1)^2} \, . \qquad \textbf{(5.2)}$$

Where r is the number of categories and i and j are the respective row and column ranks. Squared weights have an appealing asymptotic relationship with the Intra Class Correlation (ICC) (Fleiss and Cohen, 1973).

Limitations of the weighted kappa statistic have been previously cited. Graham and Jackson argued that the weighted kappa statistic is in fact not a measure of agreement but a measure of association (Graham and Jackson, 1993). They illustrated this using a hypothetical example in which one rater systematically disagrees with the other for most levels of some ordinal measure by one level yet this still yields a large weighted kappa statistic. Such behavior is undesirable for a measure of agreement. Additionally, it is highlighted that a single measure cannot possibly summarise all the information regarding agreement and disagreement. The quasi-association modelling approach outlined by Agresti may be more informative (Agresti, 2002). Despite these drawbacks, the weighted kappa is used in this chapter largely for its simplicity. It has an intuitive interpretation that is easily understood. It can be viewed as playing a similar role as the sensitivity and specificity measures in the binary case. Aspects that could possibly explain disagreement are explored and are discussed below.

5.3.2.2 The ordinal *c*-index (ORC)

Discrimination on an ordinal outcome should take the added complexity of the scale into account. The ordinal *c*-index (ORC) is a set-based measure which summarises the closeness of the predicted ordering to the observed (Van Calster et al., 2012). For example, take the OHS, which describes disability and death on a seven level ordinal scale. Suppose a complete set is observed, i.e., seven patients taking on each of the possible seven levels of OHS. A given model would ideally rank this set of seven patients in the correct order of severity, the so called 'target ordering' (i.e., the patient with outcome O_0 should have a lower predicted risk or score than the patient with outcome O_1 who should also have a lower predicted risk or score than the patient with outcome O_2 and so on up to O_6). The closeness of the predicted ordering to the target ordering is indicative of the degree of discrimination. The number of steps, S, required to put a set of predicted risks into the correct order can be made for all possible sets, with the number of proportionate steps calculated as

$$C^{steps}_{n_1,\ldots,n_k} = 1 - \frac{S}{k(k-1)/2},$$

where *k* denotes the number of category levels. When tied risks occur (i.e., the predicted risk for one patient in a set is equal to that of another), 0.5 is added to *S* for each pairwise tie. The ORC can then be interpreted as the average proportionate closeness of a set of cases to the target ranking:

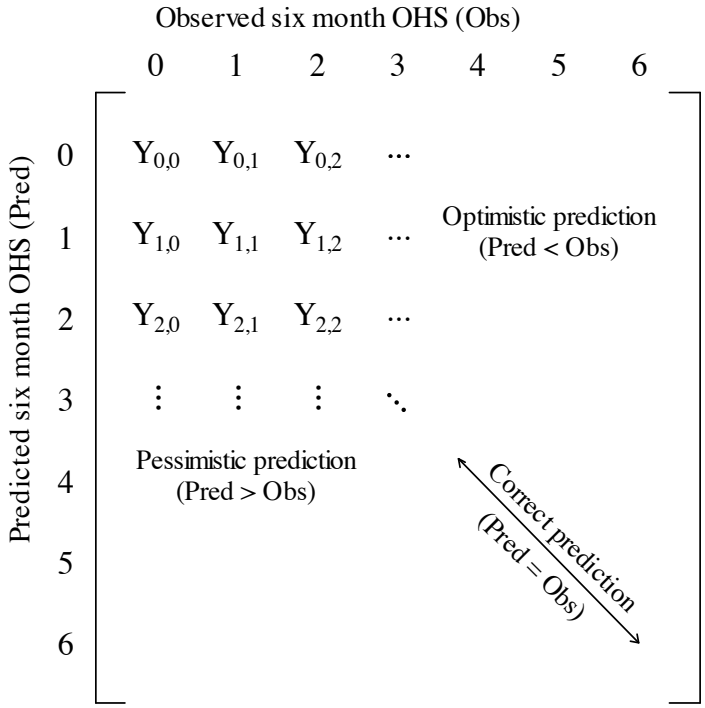
$$ORC = \frac{1}{N_1 \cdots N_k} \sum_{n_1=1}^{N_1} \cdots \sum_{n_k=1}^{N_k} C_{n_1,\ldots,n_k}^{steps} \quad . \qquad \textbf{(5.3)}$$

This measure also has a simpler pairs-based interpretation as the unweighted average of the pairwise AUROCCs. Bootstrap methods are used to calculate the 95% CI (Van Calster et al., 2012).

For each model, the linear predictor was fitted as the only predictor in a proportional odds logistic regression (POLR) for observed six month OHS. The resulting linear predictor was used to explore ordinal discrimination. To enable identifiable parameter estimates and no empty cells, the ordinal outcome was interpreted as a five level outcome, collapsing levels 4 to 6 leaving 0, 1, 2, 3 and "4 to 6" as distinct.

5.4 Methods – doctors' characteristics

Each doctor gave an informal prediction of six month OHS across the full ordered range (i.e., 0 to 6) for their own patients on initial assessment. Once the actual six month patient outcomes were observed the following classification was used as illustrated in the matrix below: (i) *correct* – when the predicted matched the observed, i.e., on the diagonal; (ii) *optimistic* – when the predicted rank was less than the observed, i.e., upper triangle; and (iii) *pessimistic* – when the predicted rank was greater than the observed, i.e., lower triangle.



There is no implicit ordering for these three classification types as there may be unknown harmful effects from the treatments administered or withheld based on these informal predictions. Multinomial logistic regression was used to explore whether: type of doctor; patient characteristics; or hospital setting were jointly influential in the classification of patients.

5.4.1 Multinomial logistic regression

Multinomial (or polytomous) logistic regression can be applied when an outcome has more than two discrete responses with no implicit ordering (Biesheuvel et al., 2008). In the general case, let outcome Y take on any value k with $k = 1, 2, \ldots, m, m + 1$. One level is set as a reference category, e.g., $m + 1$, on which all other levels are compared (Bender and Grouven, 1997). This gives m sub-models:

$$P(Y = m+1) = \frac{1}{1+\sum_{k=1}^{m}\exp(\alpha_k + \sum_{j=1}^{p}\beta_{kj}X_j)} \text{ and} \qquad \textbf{(5.4)}$$

$$P(Y = k) = \frac{\exp(\alpha_k + \sum_{j=1}^{p}\beta_{kj}X_j)}{1+\sum_{k=1}^{m}\exp(\alpha_k + \sum_{j=1}^{p}\beta_{kj}X_j)}, \text{ for } k = 1, 2, \ldots, \text{m.}$$

The scenario described in the matrix above has three unordered outcome levels (optimistic, correct and pessimistic) implying two sub-models. Two sets of measures of association each of length p (for p unspecified predictors) can be interpreted relative to the reference category. The reference category was chosen as *correct classification*, i.e., were doctors gave a predicted OHS score which matched that observed in follow-up. It therefore follows that there are two generalised logits describing two models,

$$\log\left(\frac{P(Y = optimistic)}{P(Y = correct)}\right) = \alpha_{optimistic} + \beta_1^{optimisitic}X_1 + \ldots + \beta_p^{optimisitic}X_p \text{ and}$$

$$\log\left(\frac{P(Y = pessimistic)}{P(Y = correct)}\right) = \alpha_{pessimistic} + \beta_1^{pessimistic}X_1 + \ldots + \beta_p^{pessimistic}X_p .$$

By simultaneously modelling the probability of the potential classifications, it is possible to explore the associations demonstrated in this observational cohort of stroke patients across a number of observed factors which may be influential in classification.

5.5 Results

A total of 931 patients were analysed. The flow diagram (Figure 5-2) summarises which patients in the Edinburgh Stroke Study were analysed. Of 1257 patients, 1051 had record of doctor's predictions of which 931 had complete follow-up by six months. Record of baseline measurements taken from patients on entry to the study were largely complete (Table 5-3) with low rates of missing data on those variables used by the pre-existing prediction models for functional outcome (Table 5-1).

These data are limited by consent bias. Of those that were eligible, 88% consented for their data to be part of a repository for further research, the main barrier to which was obtaining informed consent (Jackson et al., 2008). Despite this, characteristics were similar between those consenting and non-consenting patients.

To reduce the risk of bias from complete cases analyses, missing baseline data were imputed generating 20 datasets (Vergouwe et al., 2010). All of the measures listed in Table 5-3 were adopted in the imputation model. Whilst this included six month OHS all patients with missing outcome entries were removed from each imputed dataset since retaining imputed outcomes only adds random noise to the pooled results (Von Hippel, 2007).

On average those with missing six month outcomes or missing doctor's predictions were younger (median 73 years versus 74 years, P-value = 0.0130) and had less severe strokes (median NIHSS of 1 versus 2, P-value = 0.0193) (see Table 5-5 in section 5.7.1 Appendix A on page 128).

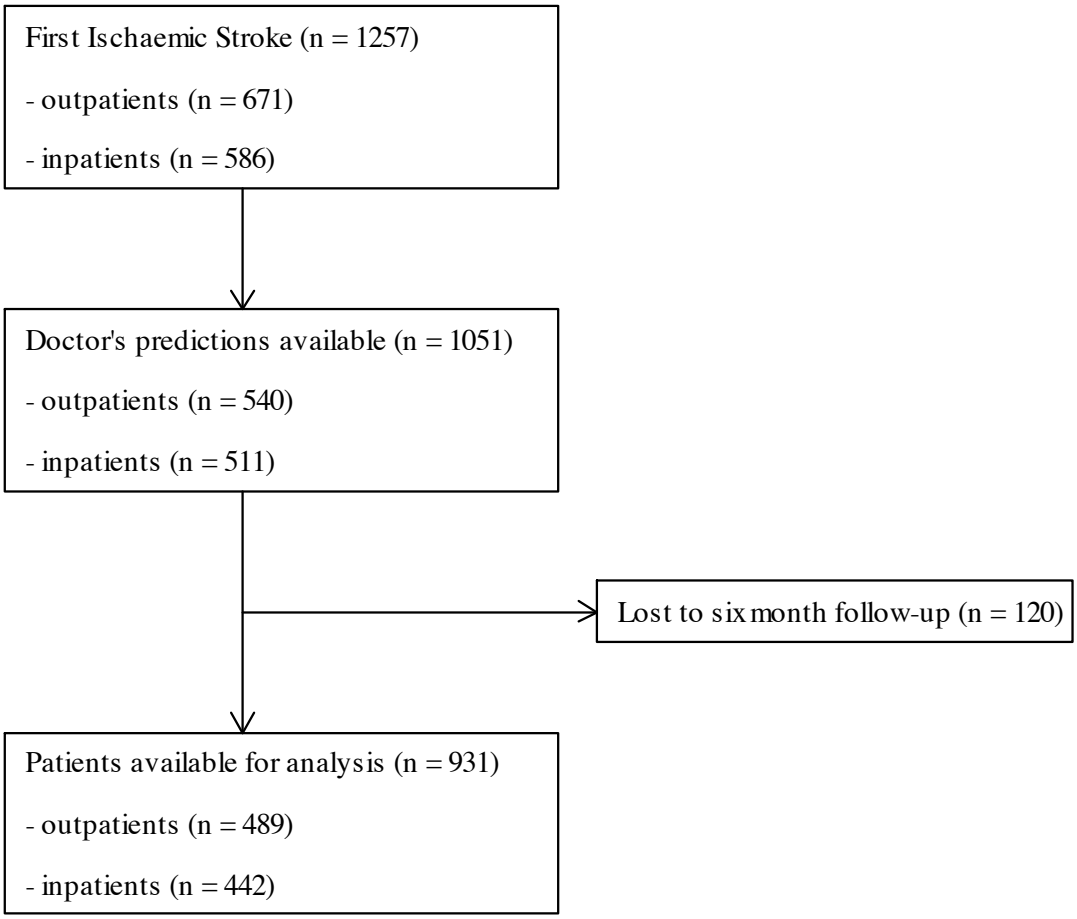


Figure 5-2 Flowchart of data available for analysis in the Edinburgh Stroke Study

Table 5-3 Characteristics of 931 ischaemic stroke patients observed in the ESS

Variable	Data (N = 931)	Number (%) missing
Doctor's clinical experience		
Fully trained vs. in training, n (%)	499 (54)	107 (11)
Stroke specialist vs. neurologist, n (%)	550 (59)	107 (11)
Baseline characteristics		
Age, years, median (IQR)	74 (66 to 81)	-
Male, n (%)	474 (51)	-
History of hypertension, n (%)	520 (56)	1 (<1)
History of diabetes mellitus, n (%)	119 (13)	-
Pre-stroke independence, n (%)	867 (93)	2 (<1)
Lived alone prior to stroke, n (%)	361 (39)	-
Arm power, n (%)	799 (86)	1 (<1)
Able to walk, n (%)	672 (72)	2 (<1)
Normal GCS verbal, n (%)	810 (87)	5 (<1)
NIHSS (median, IQR)	2 (0 to 5)	35 (4)
Heart failure, n (%)	55 (6)	2 (<1)
Total cholesterol, mmol/l, median (IQR)	5 (4 to 6)	73 (8)
Systolic BP, mmHg, median (IQR)	146 (130 to 160)	2 (<1)
Seen at outpatients, n (%)	489 (53)	-
Six month OHS score, n (%)		
0 (Fully recovered)	168 (18)	-
1	252 (27)	-
2	183 (20)	-
3	126 (14)	-
4	49 (5)	-
5	55 (6)	-
6 (Dead)	98 (11)	-

ABBREVIATIONS: Oxford Handicap Scale (OHS); National Institutes of Health Stroke Scale (NIHSS); Glasgow Comma Scale (GCS); Inter Quartile Range (IQR); and Blood Pressure (BP)

5.5.1 Model performance in ESS

The models described in Table 5-1 were applied to the Edinburgh Stroke Study data to assess their performance in an external dataset. Two aspects of model performance were assessed: discrimination and calibration. Discrimination was measured using the area under the receiver operating characteristic curve (AUROCC) and calibration was measured by calculating the associated slope and intercept estimates from the calibration plot.

5.5.1.1 Discrimination

Statistical models discriminated patients with poor outcome (OHS≥3) from those with good outcome (OHS<3) after stroke moderately well with AUROCC values ranging from 0.76 to 0.84 depending on which model was used (Table 5-4). The model developed by Reid *et al*. achieved the largest AUROCC, although there was considerable overlap between the 95% CIs of the other models (Reid et al., 2010). The model developed by Lee *et al*. achieved the lowest AUROCC (Lee et al., 2009). Whilst there were large quantitative differences noted during the sensitivity analyses this was met with only small qualitative differences. There were no important changes in the ordering of the models by AUROCC either with a poor outcome defined as OHS≥2 or OHS≥3 or when restricting to those seen in hospital or those seen at the outpatients (see section 5.7 Appendix A, Table 5-6 and Table 5-7). The largest AUROCCs were obtained when analysing all patients with poor functional outcome defined as an OHS≥3.

5.5.1.2 Calibration

Calibration was poor amongst those models that could be assessed (Table 5-4). Each systematically overestimated the risk of poor outcome except for the SSV which underestimated the risk. All models had calibration slopes less than one. This occurs when the effect estimates are on average smaller (i.e., from overfitting in development or because of true differences between the two development and evaluation cohorts) resulting in extreme risk estimates made for the ESS patients. The Weimar *et al*. model had a calibration slope of almost 1; it is likely that this model would only require slight updating to improve its calibration-in-the-large within this single centre study (Janssen et al., 2009, Weimar et al., 2004).

No improvement in calibration was found when limiting the analysis to hospital inpatients and outpatients (see section 5.7 Appendix B, Table 5-6 and Table 5-7).

5.5.2 Assessing accuracy on a dichotomous outcome

Eighteen doctors made clinical predictions: ten neurologists (56%) and eight stroke physicians (44%). Ten were in training (56%) and eight were fully trained (44%). Doctors correctly predicted level of disability or death in 310 patients (33%). A doctor's informal prediction of poor functional outcome (OHS≥3) six months after stroke had a sensitivity of 0.44 (95% 0.39 to 0.49) and a specificity of 0.96 (95% 0.94 to 0.97). The performance of clinical prediction models was similar: at the specificity of a doctor, the sensitivity of risk prediction rules to predict poor functional outcome ranged from 0.38 to 0.45; at the sensitivity of a doctor specificity of risk prediction rules ranged from 0.94 to 0.96 (Table 5-4).

There were no important differences in these results when restricting to inpatients and outpatients exclusively nor when defining six month outcome as OHS≥2 and restricting to inpatients. However, there was an indication that doctors were more accurate at identifying outpatients with OHS≥2 at six months compared to models (see section 5.7 Appendix A, Table 5-8).

5.5.3 Assessing accuracy on an ordinal outcome

Doctors predictions of OHS at six months had moderate agreement with the observed six month OHS for inpatients (weighted κ of 0.53 with 95% CI 0.42 to 0.63 on 442 patients) but had poor agreement for outpatients (0.30 with 95% CI 0.21 to 0.39 on 489 patients). Doctors tended toward optimistic prediction with 61% (95% CI 55% to 68%) of inpatients and 45% (95% CI 38% to 51%) of outpatients given a lower predicted OHS than observed at six months.

Smoothed non-parametric plots of the distribution of the linear predictor illustrate the discrimination achieved by a model since in general: the greater the spread; the greater the discrimination. The kernel density plots are provided in Figure 5-3 and show the spread of predictions made by each CPM compared to the predictions made by the doctors. There is considerable overlap in the spread of these predictions

irrespective of the method of prediction suggesting little difference in discriminative ability between formal and informal methods.

Ordinal discrimination by doctors was moderate (ORC of 0.74 with 95% CI 0.72 to 0.76) and comparable to that obtained by statistical prediction, where the probability of correctly separating a pair of patients from two randomly selected levels of the OHS (defined on categories: 0, 1, 2, 3, and ≥4) could be as low as 0.69 and as high as 0.75 (Table 5-4). Similar results applied when restricting to inpatients alone and lower discrimination when restricting to outpatients (see section 5.7 Appendix A, Table 5-9).

Clinical prediction models therefore did no better than doctors' informal prediction in ordinal discrimination.

Table 5-4 Comparison of prediction methods (Note that all intervals are 95% CIs)

	Discrimination		Calibration		Accuracy	
Method	**AUROCC[1]**	**ORC[2]**	**Intercept[1]**	**Slope[1]**	**Sensitivity[3]**	**Specificity[3]**
Doctors	-	0.74 (0.72 to 0.76)	-	-	0.44 (0.39 to 0.49)	0.96 (0.94 to 0.97)
Clinical Prediction Model						
Reid	0.84 (0.81 to 0.87)	0.75 (0.73 to 0.77)	1.09 (0.98 to 1.19)	0.46 (0.43 to 0.50)	0.45 (0.34 to 0.52)	0.96 (0.93 to 0.98)
Weimar	0.82 (0.79 to 0.85)	0.73 (0.71 to 0.76)	0.57 (0.49 to 0.65)	0.98 (0.90 to 1.06)	0.43 (0.35 to 0.51)	0.96 (0.92 to 0.98)
SSV	0.81 (0.78 to 0.84)	0.72 (0.70 to 0.74)	-0.30 (-0.39 to 0.21)	0.71 (0.66 to 0.76)	0.43 (0.36 to 0.51)	0.95 (0.93 to 0.98)
Appelros	0.82 (0.79 to 0.85)	0.73 (0.71 to 0.75)	-	-	0.42 (0.35 to 0.50)	0.95 (0.93 to 0.97)
Lee	0.76 (0.72 to 0.79)	0.69 (0.66 to 0.71)	-	-	0.38 (0.32 to 0.45)	0.94 (0.91 to 0.96)

1 Values pooled across 20 imputed datasets for missing data
2 Single imputed set, with 95% confidence intervals calculated on 1000 bootstrap replicates
3 Based on a single imputation

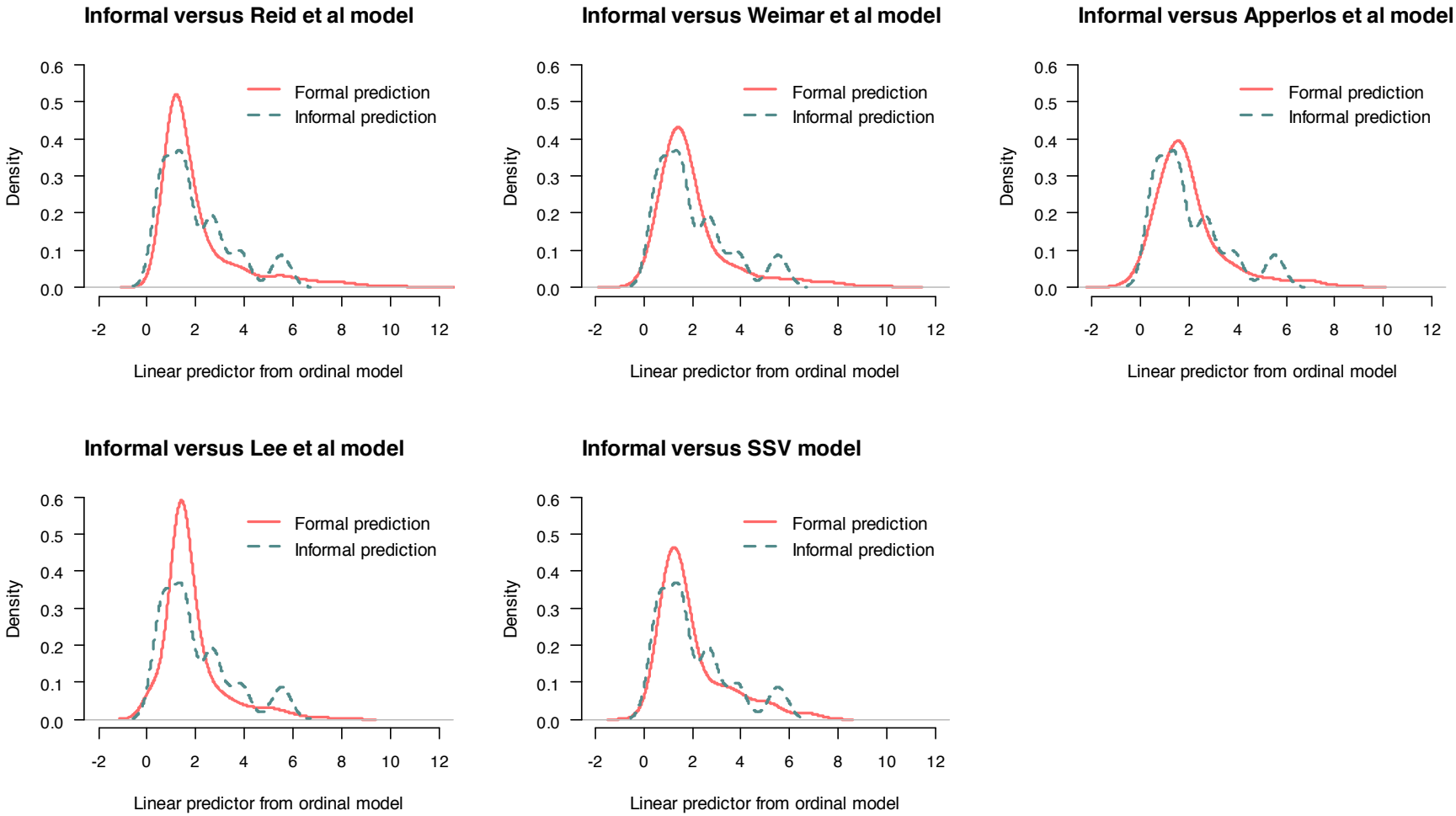


Figure 5-3 Smoothed kernel density plots of linear predictors from POLR models, comparing formal to informal prediction.

5.5.4 Doctors' characteristics

In total there were 18 doctors initials identified each of whom predicted the outcome for a number of patients. The number of predictions (i.e., patients seen) per doctor varied from 2 to 204 with a median of 30 and interquartile range 16 to 48 (Figure 5-4).

Classification was not influenced by doctor (i.e., level of training or specialty) or patient characteristics (i.e., neurological impairment, age or additional risk factors) (Figure 5-5). However, patients seen as outpatients were more likely to have had a correct prediction than an optimistic prediction of their eventual disability than those seen as inpatients (OR of 0.59 with 95% CI 0.38 to 0.92). This analysis was restricted to those patients for whom doctors' characteristics could be obtained (i.e., the 824 patients in Figure 5-4, of which: 282 (34%) where correctly classified; 418 (51%) optimistically classified; and 124 (15%) pessimistically classified). The inclusion of clinician characteristics in the model was not significant (χ^2 of 8.9 on 6 df, with a P-value = 0.1819) suggesting that no further information on classification was gained by accounting for clinicians.

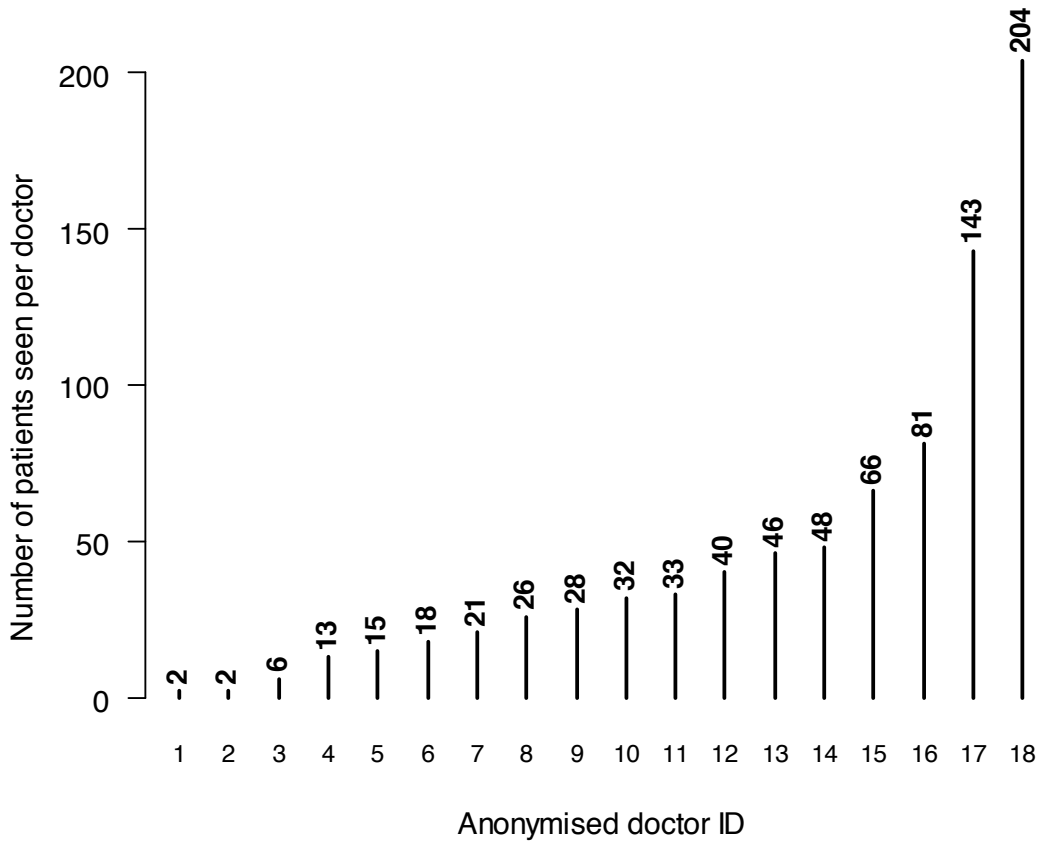


Figure 5-4 Number of patients with informally predicted outcomes per doctor.

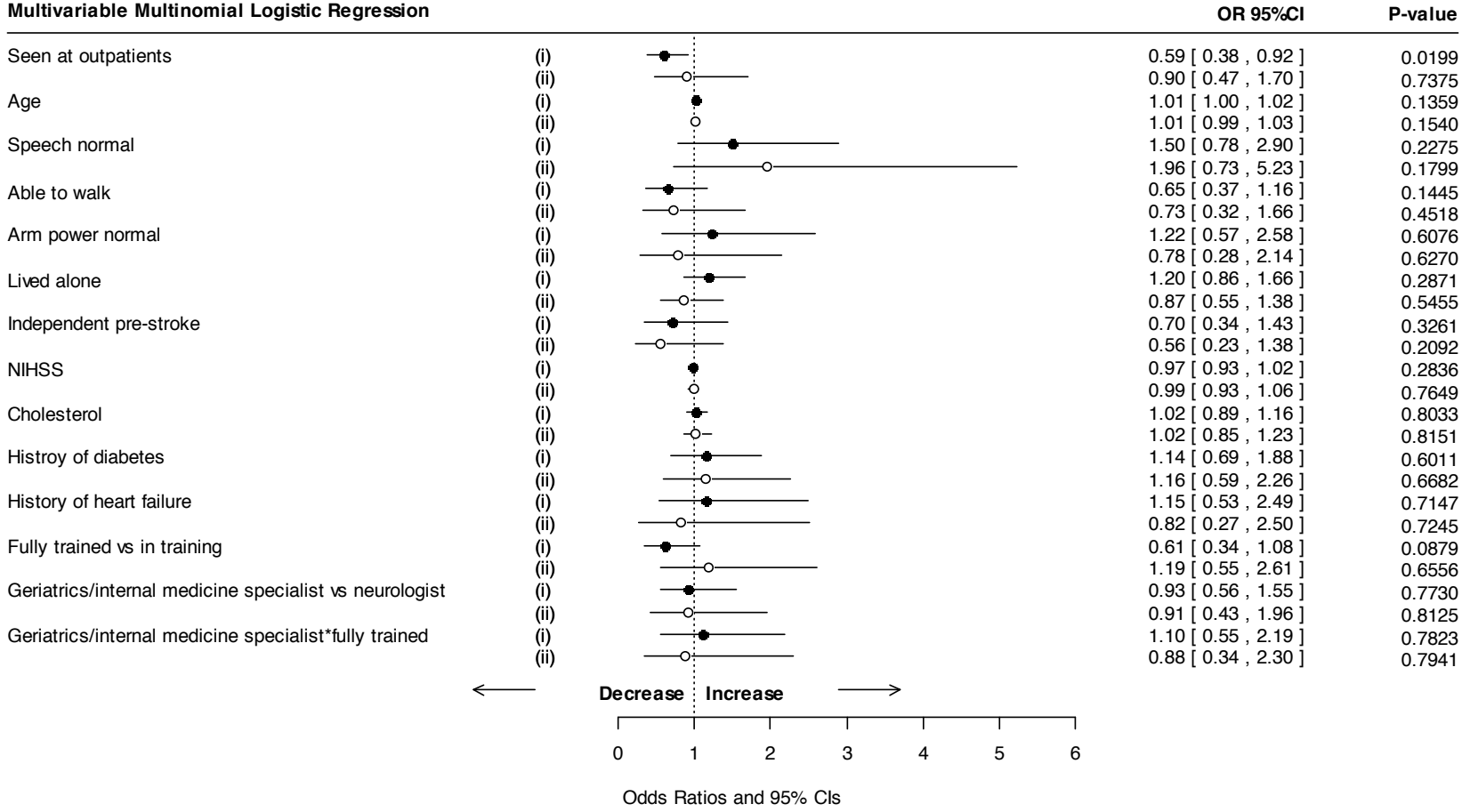


Figure 5-5 Multivariable multinomial logistic regression classified on a single imputed set. Solid points (i) denote the classifications 'optimistic vs. correct' and open points (ii) denote 'pessimistic vs. correct'.

5.6 Discussion

The accuracy of predictions made by doctors and those made by formal models were found to be similar in this single centre prospective observational study. Doctors had good specificity, but poor sensitivity at identifying those with a poor functional outcome after ischaemic stroke, however prediction models, with the corresponding sensitivity or specificity fixed at that of a clinician, achieved a similar level of accuracy.

Neither doctor characteristics nor patient characteristics made a great deal of difference to the accuracy of informal prediction; however predictions were more likely to be correct in outpatients. This is likely due to the delay in time of onset to assessment (median of 18 days with IQR 12 to 28) by which point most patients suffering minor strokes would have recovered making their observed disability at outpatients a good surrogate for their likely disability by six month. Outpatients can therefore be regarded as *closer* to their end-state in contrast to inpatients. The association between an early indication of recovery and longer term outcome has been noted elsewhere and was utilised in the third International Stroke Trial (IST-3) to impute missing six month functional outcome entries using measurements recorded at seven days within a simple algorithm (The IST-3 collaborative group, 2012b, The IST-3 collaborative group, 2012a).

More recently developed clinical prediction models, such as the ASTRAL score, the iScore and the BOAS score, could not be assessed using these data as not all of the baseline predictors were available in the ESS (Ntaios et al., 2012, Saposnik et al., 2011, Muscari et al., 2011). These models either required different measures of deficits caused by the stroke (e.g., the use of the Canadian Neurological Scale) or else a record of various other co-morbidities (e.g., cancer or renal dialysis). For the most part these characteristics were well represented by those models assessed in this chapter (Table 5-1). The level of performance demonstrated by these untested models is similar to those that were tested in this chapter (Ntaios et al., 2012, Muscari et al., 2011, Papavasileiou et al., 2013, Zhang et al., 2013). It is unlikely that the inclusion of these variables would result in any considerable improvement

making the findings in this chapter broadly generalisable to any closely related models.

Some of the prediction models tested in this chapter were developed to predict distinct outcomes or used predictors with different definitions. Specifically, the model developed by Reid *et al*. included a stroke severity score which – though useful – is rarely used in practice (The EC/IC Bypass Study group, 1985). The NIHSS was used in its place where it was found that regardless of any qualitative differences between these two measures the Reid model performed well. This is likely due to a strong correlation between the stroke severity score and the NIHSS (Goldie et al., 2014). This is therefore supportive of a good degree of generalisability in the discriminatory ability of these models. Updating would likely be required to improve upon calibration (Collins and Le Manach, 2013).

Similar studies have suggested conflicting findings. A simulation study demonstrated that doctors predictions were worse than model-based predictions, albeit based on scenario-based, rather than clinical predictions (Saposnik et al., 2013a). However, the findings made in this chapter support those made in a previous study which demonstrated that doctors predictions of poor functional outcome were similar to the six simple variables model (Counsell et al., 2004). Additionally, patients' functional outcome was examined in this chapter whilst maintaining its true ordinal structure which has not been investigated in such detail before.

The inclusion of more complex variables such as the NIHSS and stroke subtype may seem intuitive, though they add to the difficulty of using these models, which may limit their use by non-specialists, non-doctors, and doctors early in their training. In any case, the main distinction between the SSV model and the Reid *et al*. model is the use of a stroke severity measure (note that the NIHSS was used in this chapter). The performance of these models was largely the same in the ESS, with the suggestion that the Reid *et al*. model holds the edge on the SSV in predicting poor functional outcome in the outpatients (see section 5.7 Appendix A on page 128).

5.7 Appendix A: Extra tables

5.7.1 Loss to follow-up and missing doctor's predictions

A number of patients were either missing an OHS score at six months or missing a doctor's prediction. Such patients were excluded from all analyses. The associations at a univariate level are shown in Table 5-5.

5.7.2 Sensitivity analyses

Table 5-6, Table 5-7 Table 5-8 and Table 5-9 summarise the relevant analyses presented above in the Results section within subsets and differently defined patients of the Edinburgh Stroke Study.

Table 5-5 Prevalence of risk factors at baseline in those included in analysis vs. those with either missing informal prediction or missing observed outcome at six month follow-up. Data presented as number and percentage (%) unless otherwise stated.

	Patients available for analysis (n = 931)		Missing outcome or prediction (n = 326)		
Measurements taken on entry	**No. (%)**	**No. Missing**	**No. (%)**	**No. Missing**	**P-value**
Variables used in formal prediction models					
Age (years) (median, IQR)	74 (66 to 81)	-	73 (61 to 81)	-	0.0130
Pre-stroke independence	867 (93)	2	298 (91)	3	0.5167
Lived alone prior to stroke	361 (39)	-	120 (37)	2	0.5794
Arm power	799 (86)	1	276 (85)	2	0.7468
Able to walk	672 (72)	2	240 (74)	4	0.4445
Normal GCS verbal	810 (87)	5	279 (86)	3	0.6121
NIHSS (median, IQR)	2 (0 to 5)	35	1 (0 to 3)	42	0.0193
Heart failure	55 (6)	2	25 (8)	1	0.2619
History of diabetes	119 (13)	-	39 (12)	-	0.7012
Total cholesterol (mmol/l) (median, IQR)	5 (4 to 6)	73	5 (4 to 6)	46	0.3635
Additional variables					
Systolic BP (mmHg) (median, IQR)	146 (130 to 160)	2	148 (130 to 164)	5	0.9778
Gender, Male	474 (51)	-	170 (52)	-	0.7012
History of hypertension	520 (56)	1	163 (50)	-	0.0653
Atrial fibrillation	66 (20)	2	205 (22)	-	0.4918
Able to talk	844 (91)	1	295 (90)	2	0.8734

Table 5-5 Continued from previous page

	Patients available for analysis (n = 931)		Missing outcome or prediction (n = 326)		
Measurements taken on entry	**No. (%)**	**No. Missing**	**No. (%)**	**No. Missing**	**P-value**
Additional variables (continued)					
Stroke syndrome					
TACS	97 (11)		32 (10)		-
LACS	250 (28)		73 (24)		-
PACS	412 (47)		149 (49)		-
POCS	122 (14)		53 (17)		-
Missing	-	50	-	19	0.2853
Seen at outpatients	489 (53)	-	182 (56)	-	0.3036

ABBREVIATIONS: Total Anterior Circulation Stroke (TACS), Lacunar stroke (LACS), Partial Anterior Circulation Stroke (PACS), and Posterior Circulation Stroke (POCS)

Table 5-6 Model performance split by whether patients were seen in hospital or in outpatients. Poor functional outcome defined as OHS≥3

			Calibration		Discrimination
Patients	**Median R^2 (%, IQR)**	**AIC**	**Intercept (95%CI)**	**Slope (95%CI)**	**AUROCC (95%CI)**
Inpatients					
Reid	34.7 (34.4 to 34.9)	204	0.84 (0.69 to 0.98)	0.35 (0.31 to 0.39)	0.82 (0.78 to 0.86)
Weimar	36.5 (36.1 to 36.8)	196	0.77 (0.65 to 0.88)	0.79 (0.70 to 0.89)	0.82 (0.78 to 0.86)
SSV	37.6 (37.3 to 37.7)	199	-0.11 (-0.24 to 0.02)	0.61 (0.55 to 0.68)	0.82 (0.78 to 0.86)
Apperlos*	37.6 (37.4 to 37.8)	197	-	-	0.81 (0.77 to 0.85)
Lee*	24.6 (24.3 to 24.7)	130	-	-	0.75 (0.70 to 0.79)
Outpatients					
Reid	12.0 (11.7 to 12.1)	66	1.33 (1.19 to 1.46)	0.68 (0.58 to 0.79)	0.71 (0.65 to 0.78)
Weimar	8.8 (8.6 to 9.2)	56	0.36 (0.24 to 0.48)	1.04 (0.82 to 1.26)	0.67 (0.60 to 0.73)
SSV	8.6 (8.6 to 8.6)	62	-0.48 (-0.61 to -0.35)	0.54 (0.43 to 0.66)	0.64 (0.57 to 0.71)
Apperlos*	7.0 (6.8 to 7.3)	53	-	-	0.66 (0.59 to 0.73)
Lee*	0.0 (0.0 to 0.1)	24	-	-	0.57 (0.50 to 0.65)

Table 5-6 Continued from previous page

Patients	Median R^2 (%, IQR)	AIC	Calibration		Discrimination
			Intercept (95%CI)	Slope (95%CI)	AUROCC (95%CI)
All patients					
Reid	41.8 (41.6 to 42.0)	334	1.09 (0.98 to 1.19)	0.46 (0.43 to 0.50)	0.84 (0.81 to 0.87)
Weimar	40.0 (39.9 to 40.2)	317	0.57 (0.49 to 0.65)	0.98 (0.90 to 1.06)	0.82 (0.79 to 0.85)
SSV	39.2 (39.0 to 39.5)	319	-0.30 (-0.39 to -0.21)	0.71 (0.66 to 0.76)	0.81 (0.78 to 0.84)
Apperlos*	39.6 (39.5 to 39.8)	313	-	-	0.82 (0.79 to 0.85)
Lee*	27.9 (27.7 to 28.0)	207	-	-	0.76 (0.72 to 0.79)

Note: Based on a single imputation; with bootstrap 95% CIs fixed at clinicians informal sensitivity/specificity. Also, *no calibration possible as intercepts were not made available

Table 5-7 Model performance split by whether patients were seen in hospital or in outpatients. Poor functional outcome defined as OHS≥2

			Calibration		Discrimination
Patients	**Median R^2 (%, IQR)**	**AIC**	**Intercept (95%CI)**	**Slope (95%CI)**	**AUROCC (95%CI)**
Inpatients					
Reid	22.0 (21.5 to 22.3)	129	2.50 (2.36 to 2.64)	0.28 (0.24 to 0.33)	0.77 (0.72 to 0.82)
Weimar	21.5 (20.9 to 21.7)	117	1.96 (1.83 to 2.08)	0.61 (0.52 to 0.71)	0.77 (0.72 to 0.82)
SSV	24.0 (23.8 to 24.3)	133	1.34 (1.21 to 1.48)	0.53 (0.46 to 0.59)	0.78 (0.73 to 0.82)
Apperlos*	22.9 (22.5 to 23.1)	116	-	-	0.77 (0.72 to 0.82)
Lee*	13.0 (12.8 to 13.3)	69	-	-	0.70 (0.65 to 0.75)
Outpatients					
Reid	9.9 (9.6 to 10.2)	56	2.69 (2.59 to 2.79)	0.60 (0.50 to 0.70)	0.66 (0.61 to 0.71)
Weimar	6.9 (6.6 to 7.0)	45	1.52 (1.43 to 1.62)	0.73 (0.57 to 0.90)	0.63 (0.58 to 0.68)
SSV	7.8 (7.8 to 7.8)	47	0.80 (0.70 to 0.90)	0.52 (0.41 to 0.63)	0.62 (0.56 to 0.67)
Apperlos*	4.9 (4.7 to 4.9)	42	-	-	0.62 (0.57 to 0.67)
Lee*	0.11 (0.10 to 0.15)	21	-	-	0.57 (0.51 to 0.62)

Table 5-7 Continued from previous page

Patients	Median R^2 (%, IQR)	AIC	Calibration		Discrimination
			Intercept (95%CI)	Slope (95%CI)	AUROCC (95%CI)
All patients					
Reid	29.5 (29.3 to 30.0)	227	2.62 (2.54 to 2.70)	0.44 (0.40 to 0.48)	0.78 (0.75 to 0.81)
Weimar	27.1 (26.9 to 27.2)	207	1.68 (1.61 to 1.76)	0.82 (0.74 to 0.90)	0.77 (0.74 to 0.80)
SSV	26.7 (26.5 to 27.1)	221	0.99 (0.92 to 1.07)	0.64 (0.58 to 0.69)	0.75 (0.72 to 0.78)
Apperlos*	26.5 (26.4 to 26.6)	202	-	-	0.76 (0.73 to 0.79)
Lee*	17.3 (17.2 to 17.4)	125	-	-	0.71 (0.67 to 0.74)

Note: Based on a single imputation; with bootstrap 95% CIs fixed at clinicians informal sensitivity/specificity. Also, *no calibration possible as intercepts were not made available

Table 5-8 Sensitivity and specificity for formal and informal prediction methods split by where the patient was seen for two common dichotomies of OHS

	OHS≥2		OHS≥3	
Patients	**Sensitivity (95%CI)**	**Specificity (95%CI)**	**Sensitivity (95%CI)**	**Specificity (95%CI)**
Inpatients				
Reid	0.60 (0.50 to 0.73)	0.73 (0.64 to 0.83)	0.51 (0.39 to 0.60)	0.89 (0.82 to 0.94)
Weimar	0.60 (0.49 to 0.74)	0.74 (0.64 to 0.82)	0.51 (0.42 to 0.61)	0.89 (0.82 to 0.94)
SSV	0.62 (0.51 to 0.71)	0.72 (0.60 to 0.82)	0.52 (0.42 to 0.63)	0.90 (0.84 to 0.95)
Apperlos	0.61 (0.53 to 0.75)	0.74 (0.65 to 0.83)	0.52 (0.41 to 0.61)	0.89 (0.82 to 0.94)
Lee	0.54 (0.45 to 0.65)	0.60 (0.47 to 0.72)	0.47 (0.37 to 0.54)	0.84 (0.76 to 0.91)
Doctor's predictions	0.69 (0.64 to 0.74)	0.79 (0.70 to 0.85)	0.55 (0.49 to 0.62)	0.91 (0.87 to 0.95)
Outpatients				
Reid	0.28 (0.20 to 0.34)	0.86 (0.80 to 0.90)	0.16 (0.10 to 0.30)	0.99 (0.97 to 1.00)
Weimar	0.21 (0.13 to 0.29)	0.80 (0.75 to 0.86)	0.10 (0.02 to 0.20)	0.98 (0.95 to 1.00)
SSV	0.24 (0.17 to 0.30)	0.79 (0.71 to 0.85)	0.15 (0.05 to 0.27)	0.98 (0.97 to 1.00)
Apperlos	0.19 (0.11 to 0.30)	0.81 (0.72 to 0.87)	0.09 (0.01 to 0.17)	0.98 (0.94 to 0.99)
Lee	0.18 (0.11 to 0.28)	0.80 (0.68 to 0.87)	0.07 (0.01 to 0.15)	0.97 (0.93 to 0.99)
Doctor's predictions	0.38 (0.31 to 0.45)	0.92 (0.88 to 0.94)	0.09 (0.04 to 0.16)	0.98 (0.96 to 0.99)

Table 5-8 Continued from previous page

	OHS≥2		OHS≥3	
Patients	**Sensitivity (95%CI)**	**Specificity (95%CI)**	**Sensitivity (95%CI)**	**Specificity (95%CI)**
All patients				
Reid	0.56 (0.49 to 0.63)	0.87 (0.83 to 0.90)	0.45 (0.34 to 0.52)	0.96 (0.93 to 0.98)
Weimar	0.54 (0.47 to 0.60)	0.85 (0.79 to 0.90)	0.43 (0.35 to 0.51)	0.96 (0.92 to 0.98)
SSV	0.52 (0.46 to 0.58)	0.80 (0.74 to 0.89)	0.43 (0.36 to 0.51)	0.95 (0.93 to 0.98)
Apperlos	0.52 (0.44 to 0.58)	0.83 (0.78 to 0.88)	0.42 (0.35 to 0.50)	0.95 (0.93 to 0.97)
Lee	0.47 (0.41 to 0.53)	0.78 (0.72 to 0.84)	0.38 (0.32 to 0.45)	0.94 (0.91 to 0.96)
Doctor's predictions	0.58 (0.53 to 0.62)	0.88 (0.85 to 0.91)	0.44 (0.39 to 0.49)	0.96 (0.94 to 0.97)

NOTE: Based on a single imputation; with bootstrap 95% CIs fixed at doctors' informal sensitivity/specificity and 95% ZL CIs calculated for binomial proportions

Table 5-9 Performance of formal and informal prediction of the ordinal OHS (defined on five levels: 0, 1, 2, 3 and ≥4)

	Inpatients (N = 442)		Outpatients (N = 489)	
Method of prediction	**AIC**	**ORC (95%CI)**	**AIC**	**ORC (95%CI)**
Doctor's predictions	191	0.72 (0.69 to 0.77)	79	0.65 (0.61 to 0.69)
Formal: statistical model				
Reid et al	182	0.71 (0.68 to 0.75)	54	0.65 (0.61 to 0.70)
Apperlos et al	192	0.72 (0.68 to 0.76)	16	0.61 (0.57 to 0.66)
Weimar et al	185	0.71 (0.68 to 0.75)	21	0.62 (0.58 to 0.66)
SSV	190	0.72 (0.68 to 0.75)	26	0.60 (0.55 to 0.65)
Lee et al	122	0.66 (0.62 to 0.71)	-2	0.57 (0.51 to 0.62)

NOTE: A larger AIC indicates a better fit. The ORC CIs are calculated over 1000 bootstrap replicates

Chapter 6: Benefits and harms from aspirin in acute ischaemic stroke

Background and summary

Patients at a high predicted risk of further thrombosis or a low predicted risk of haemorrhage may experience greater benefit from aspirin. Additionally, absolute benefit from treatment may increase with baseline prognosis risk. This chapter explores these questions using individual patient data from three randomised trials of aspirin in acute stroke. There was no evidence to support targeting aspirin to acute ischaemic stroke patients with a high predicted risk of thrombosis or a low predicted risk of haemorrhage.

6.1 Introduction

Aspirin has been established as an effective and beneficial treatment for patients with acute stroke (Sandercock et al., 2008). Aspirin is an antiplatelet drug which reduces the chance of clot formation and subsequent thrombotic events (e.g., ischaemic stroke or myocardial infarction) (Paikin and Eikelboom, 2012, Antithrombotic-Trialists'-Collaboration, 2002). The Cochrane review of antiplatelet therapy in acute ischaemic stroke showed that aspirin was of overall benefit with 13 more patients alive and independent for every 1000 treated, i.e., an absolute risk reduction in poor functional outcome of just over 1% (Sandercock et al., 2008). An unfortunate side-effect of aspirin is that it increases patient risk of suffering intra-cranial or extra-cranial haemorrhage. One understanding of the net benefit of aspirin is that the reduction in thrombotic events out-weighs the harmful increase in haemorrhage (Bednar and Gross, 1999).

For any given trial baseline prognostic risk in the absence of treatment will vary from patient to patient. Kent and Hayward note that this risk typically has a strong positive skew with a small group of high risk patients suffering the majority of outcome

events in follow-up (Kent and Hayward, 2007b). For a constant relative risk reduction, those at a low baseline outcome risk often do not stand to gain as much benefit from treatment as those at a high risk. Absolute treatment benefit therefore necessarily varies with baseline risk. This has been called the 'heterogeneity of treatment effect' (*HTE*) (Kent et al., 2010). In the absence of treatment related harm the problem may be entirely framed as one of the cost-effectiveness of treatment. However, when treatment comes with a harmful side effect (e.g., haemorrhage); then it may be the case that the size of any potential benefit for those at low baseline outcome risk may be far less than the cost with respect to harm. The benefit harm balance may therefore tip in favour of harm more frequently amongst those least likely to benefit from treatment.

Often trials will investigate *HTE* using a plethora of subgroup analyses. Hayward *et al* demonstrated that subgroup analyses are frequently underpowered and that by stratifying by predicted risk based on multivariable prediction models it was possible to regain some of this lost power (Hayward et al., 2006). Recently Kent *et al*. proposed a formalised framework to explore *HTE* using prediction models (Kent et al., 2010). This type of approach has gained momentum with, for example, the *HTE* for various interventions in cardiovascular disease explored using existing randomised trial data and risk scores (Dorresteijn et al., 2011a, Dorresteijn et al., 2011b, van der Leeuw et al., 2014). With the associated harms and benefits of aspirin in acute stroke there is scope to explore treatment effect by predicted risk (see Chapter 1 for more discussion). The predicted risk of treatment related harm and benefit was not considered in the Cochrane review of aspirin (Sandercock et al., 2008). It is hypothesised that treating those with the greatest expected benefit (high risk of thrombosis) or those with the lowest expected harm (low risk of haemorrhage) may further reduce the risk of poor functional outcome following stroke.

The aims of this chapter are therefore twofold: (i) to develop and evaluate new clinical prediction models from large RCT data for the separate prediction of haemorrhagic, thrombotic, and poor functional outcome events; and (ii) to use these models to suggest a stratified approach for treating patients with aspirin based on identifiable predicted risk groups.

6.2 Meta-analysis and Individual Patient Data (IPD)

Three distinct *types* of meta-analysis can be identified (Senn, 2000). Using Senn's terminology: *type* A distinguishes those analyses for which the individual patient data have been made available, each recording the same outcome measure; *type* B meta-analyses are perhaps the most commonly encountered of the three using summary statistics obtained from publication, again with each trial using the same outcome measure; and finally *type* C which describes a 'unit-free' approach. This was first proposed by Glass (Glass, 1976) for summarising trials measuring efficacy of treatment on different outcomes. The latter of these approaches is irrelevant so far as any of the analyses undertaken throughout this thesis and will not be discussed any further.

International collaboration has seen the sharing of entire datasets allowing for the *type* A method to be used more frequently. This approach is more commonly known as Individual Patient Data Meta-Analysis (IPDMA) which enables a level of continuity that a traditional *type* B meta-analysis may be incapable of achieving (Debray et al., 2013). For instance, pooling measures adjusted for important (and recorded) prognostic covariates is entirely reliant on the original authors publishing these results. With IPD though detailed risk prediction analyses (i.e., model development and model evaluation) can be carried out. This can be especially useful when the outcome event rate is low; this pooling of data may give researchers the statistical power needed to reliably estimate the model parameters (Ahmed et al., 2014). In summary, individual patient data sources can be appropriately combined and used to answer important questions that smaller studies cannot achieve in isolation.

The latest Cochrane review of antiplatelet therapy for acute ischemic stroke was used to identify trials of aspirin versus placebo or control in acute ischemic stroke (Sandercock et al., 2008). It was not possible to obtain the individual patient data for two small trials one of which was unpublished (Pince, 1981, Rödén-Jüllig et al., 2003). The individual patient data obtained for analysis in this chapter comprise three

of the largest randomised trials of aspirin versus control (inactive or active) in acute ischaemic stroke (summary provided in Table 6-1).

6.2.1 The first International Stroke Trial (IST-1)

The first International Stroke Trial (IST-1) was a large, randomised, factorial designed, open control trial of up to 14 days of aspirin and two doses of heparin started within 48 hours of acute stroke onset (IST Collaborative Group, 1997). Patients were followed up for a maximum of six months. The primary outcomes were: (i) death from any cause within 14 days of randomisation and (ii) death or dependency at six months. A minimisation algorithm was used to ensure balance between the treatment arms for a number of "*…recorded prognostic features…*" although it is not clear from the original publication what exactly these were. The IST-1 had a 2x3 factorial design randomising patients to receive either: (i) 300mg/day of aspirin and 12,500 IU bd heparin per day (medium-dose); (ii) 300mg/day of aspirin and 5000 IU bd heparin per day (low-dose); (iii) 300mg/day of aspirin and no heparin; (iv) no aspirin and 12,500 IU bd heparin per day; (v) no aspirin and 5000 IU bd heparin per day; and (vi) no aspirin and no heparin. The factorial design was adopted to investigate whether any important interactions occurred between aspirin and heparin (low or medium dose). Patients were recruited from January 1991 till May 1996 from 36 countries and 467 hospitals. A total of 19435 patients were randomised to receive one of the six combinations listed above. Follow-up was good with 19285 (99.2%) of those randomised with observed outcomes by six months (a total of 150 lost, ranging from 0.9 to 0.6% across the six treatment regimens). No significant differences were identified in early deaths (≤14 days) amongst those randomised to receive either heparin (low and medium doses pooled) or no heparin or amongst those randomised to receive either aspirin or no aspirin (P-values ≥ 0.05). Similarly there was no difference in those dead or dependent at six months between treatment arms, although, after adjustment for baseline stroke severity a beneficial reduction of 14 per 1000 treated (P-value = 0.03) was identified. The trialists found that no treatment interaction existed between aspirin and heparin in their study.

6.2.2 The Chinese Acute Stroke Trial (CAST)

The Chinese Acute Stroke Trial (CAST) was a large, randomised, placebo controlled trial of up to 28 days of aspirin started within 48 hours of acute stroke onset (Zheng-Ming, 1997). Patients were followed up for a maximum of 28 days. The primary outcomes were: (i) death from any cause within 28 days; and (ii) death or dependence at discharge. The CAST recruited patients from November 1993 till March 1997 from 413 Chinese hospitals. A total of 21106 patients were randomised to receive either 160mg/day aspirin or an inactive placebo. Follow-up was good with 20655 (97.9%) of those randomised with available discharge forms by 28 days (a total of 451 lost, with 2.1% in the aspirin arm and 2.4% in placebo). A significant treatment effect was identified amongst those dead by 28 days, with around 6 fewer deaths per 1000 treated with aspirin (P-value = 0.03). Of those dead or dependent at discharge there were around 11 per 1000 fewer (P-value = 0.08).

6.2.3 Multicentre Acute Stroke Trial – Italy (MAST-I)

The Multicentre Acute Stroke Trial – Italy (MAST-I) was a small, randomised, factorial designed trial of streptokinase or aspirin (Multicentre Acute Stroke Trial - Italy, 1995). The primary outcome was death or severe disability six months from randomisation. The purpose of the trial was to assess whether there was any benefit from aspirin and/or streptokinase in reducing the incidence of death or severe disability. The MAST-I had a 2x2 factorial design randomising patients to receive either: (i) a one hour intravenous infusion of 1.5 MU streptokinase; (ii) 300mg/day of buffered aspirin for ten days; (iii) both active treatments; or (iv) neither. Four interim analyses were planned after the recruitment of 100, 300, 500 and 600 patients. The MAST-I recruited 622 patients between May 1991 and February 1995 out of a planned total of 1500 patients. No patients were lost to follow-up. The trial was interrupted in January 1995 due to an excess number of early deaths amongst those randomised to receive streptokinase. A significant excess of deaths by 10 days was identified amongst those randomised to receive either of the streptokinase regimens. By six months a non-significant reduction in those dead or severely disabled was suggested for both those randomised to receive streptokinase (OR 0.9 with 95% CI 0.7 to 1.3) and for those on aspirin (OR 0.9 with 95% CI 0.6 to 1.3).

Table 6-1 Characteristics of included aspirin trials

Design Features	IST-1	CAST	MAST-I
Number of patients	19435	21106	622
Aspirin dose, mg	300	160	300
Duration, days	14	28	10
Non-aspirin treatment arm(s)	Low dose/ medium dose/ or avoid heparin	Placebo	Streptokinase or avoid streptokinase
Randomisation	2x3 factorial	1:1	2x2 factorial
Time from onset, hours	0 to 48	0 to 48	0 to 6
Follow up for early events	14 days	28 days	10 days
Follow up for death or dependence	Six months	28 days	Six months

6.2.4 Patient characteristics across trials

A total of 39166 patients with acute ischemic stroke from the three trials were suitable for this analysis. Patients with definite or probable (i.e., unknown pathology) ischaemic stroke at baseline were retained.

6.2.4.1 Baseline covariate data

Most baseline characteristics were used as originally defined or were modified so as to induce a degree of communality across the three trials. MAST-I recorded the presence of a weakness in the 'arm or leg' as a single binary variable, this variable was therefore defined within both IST-1 and CAST and adopted throughout all analyses. Age was parameterised per increasing intervals of ten years and systolic blood pressure per increasing intervals of ten millimetres of mercury (mmHg). The remaining variables were used as originally measured: presence of dysphasia, hemianopia, visuospatial disorder, brainstem or cerebellar deficits, a history of atrial fibrillation, delay from stroke to randomisation, and evidence of infarct on CT. In order to assess the modest but beneficial effect of aspirin both the IST-1 and CAST trials were designed pragmatically. Pragmatic trials help to evaluate the applicability of primary findings in a broader context and ultimately have a greater influence on health services and policy making (Treweek and Zwarenstein, 2009). However, the pragmatism employed in the running of these trials meant that few baseline measurements were taken thus placing a necessary limit on the amount of variable selection possible in this secondary-analysis of the data.

Patients in CAST were on average younger (median age of 64); had a lower SBP (median of 154mmHg) and were more often male (63%) in contrast with those recruited to IST-1 and MAST (similar median age (73 and 71 respectively), equal SBP (160mmHg) and an equal proportion who were male (54%), see Table 6-3). Stroke deficits were largely comparable between the three trials with the exception of hemianopia and visuospatial deficit which occurred less often and dysphasia which occurred more often in CAST than in IST-1 and MAST.

6.2.4.2 Early and late outcomes

Two early composite events were defined: a thrombotic event (any Deep Venous Thrombosis (DVT); Pulmonary Embolism (PE); ischaemic stroke; and Myocardial Infarction (MI)) and a haemorrhagic event (any significant intracranial haemorrhage; major extra-cranial haemorrhage; and haemorrhagic transformation). IST-1 reported these events at 14 days. In a sensitivity analysis those with DVTs were excluded (see section 6.6). Early events in CAST were restricted up to and including 14 days whilst MAST-I recorded early events at 10 days. Differences in incidence were noted across trials with: 6.1% of patients suffering a thrombotic event and 1.4% a haemorrhagic event in IST-1; 1.9% and 1.0% in CAST; and 1.6% and 5.8% in MAST-I. A particularly high rate of haemorrhage in MAST-I can be attributed to an excess risk amongst those treated with streptokinase: this was the reason for the early termination of MAST-I (Multicentre Acute Stroke Trial - Italy, 1995). Allowances for these differences were made through recalibrated predicted risks which will be discussed in detail later (see section 6.4.1).

Each trial used slightly different ordinal scales to record functional outcome by end of follow-up. The IST-1 recorded functional outcome at six months. This ordered scale ranged from 1 to 4, describing patients as: *dead*; *dependent*; *not recovered*; and *recovered*. The original publication for CAST claims a record of functional outcome on the modified Rankin Scale (mRS), although on inspection of the available IPD a measure similar to that used in IST-1 seems to have been recorded at 28 days describing patients as: *dead*; *partly dependent*; *severely disabled*; *recovered, but independent*; and *fully recovered*. Finally, MAST-I assessed patients using a more conventional ordinal scale by six months: the modified Rankin Score (mRS). It was possible to define a 'common ordinal outcome' across the three trials. As IST-1 offered the most restricted number of levels the ordinal outcomes for CAST and MAST-I were constrained and categorised in order to fit within the constructs of the four levelled IST-1 outcome scale. Therefore an assumption that patients so defined are comparable across trials is made. This *common ordinal outcome* is summarised below (Table 6-2). Poor functional outcome was defined as those dead or dependent, i.e., a score of 2 or less on the common ordinal outcome scale. Almost two-thirds of

patients in IST-1 (62%) and MAST-I (64%) had a poor functional outcome by six months compared to just under a third of those in CAST (31%) by 28 days (Table 6-4).

Table 6-2 Defined common ordinal outcome

	Ordinal category level (1 to 4)			
Trial	**1**	**2**	**3**	**4**
IST-1	Dead	Dependent	Not recovered	Recovered
CAST	Dead	Partly dependent OR severely disabled	Partly recovered, but independent	Fully recovered
MAST-I	mRS = 6	mRS = 3 to 5	mRS = 2	mRS = 0 to 1

Abbreviations: modified Rankin Scale - mRS

Table 6-3 Baseline characteristics for the three aspirin trials. For continuous measurements the median and inter-quartile range is quoted; for categorical measurements the frequency and % is provided.

Variable	IST-1 (N = 18372)	Missing	CAST (N = 20172)	Missing	MAST-I (N = 622)	Missing
Age (per year)	73 (65-80)	-	64 (57-70)	66 (<1)	71 (62-78)	-
Gender, male	9855 (54)	-	12759 (63)	-	335 (54)	-
Time from onset, hours	20 (9-29)		24 (12-37)	25 (<1)	4 (3-5)	-
Prior antiplatelet	729 (4)	13 (<1)	1590 (8)	534 (3)	-	-
Stroke syndrome		52 (<1)	-	-	-	-
Total anterior	4395 (24)	-	-	-	-	-
Partial anterior	7395 (40)	-	-	-	-	-
Lacunar circulation	4428 (24)	-	-	-	-	-
Posterior circulation	2103 (11)	-	-	-	-	-
Systolic Blood Pressure (mmHg)	160 (140-180)	-	154 (140-180)	19 (<1)	160 (140-170)	3 (<1)
Presence of Deficits						
Face	13392 (73)	231 (1)	14999 (74)	484 (2)	-	-
Arm/hand	15781 (86)	114 (1)	17958 (89)	161 (1)	-	-
Leg/foot	13899 (76)	238 (1)	17872 (89)	168 (1)	-	-
Either arm/leg	15920 (87)	258 (1)	18505 (92)	222 (1)	527 (85)	-
Dysphasia	8041 (44)	554 (3)	10792 (53)	322 (2)	265 (43)	-
Hemianopia	2924 (16)	3712 (20)	915 (5)	1986 (10)	71 (11)	-
Visuospatial	2992 (16)	3255 (18)	722 (4)	1997 (10)	82 (13)	-
Brainstem/cerebellar signs	2018 (11)	1479 (8)	1350 (7)	1135 (6)	16 (3)	-
Other	1143 (6)	1165 (6)	1370 (7)	1645 (8)	-	-
Conscious at randomisation	14117 (77)	-	17588 (87)	24 (<1)	568 (91)	-
Atrial Fibrillation (AF)	3034 (17)	947 (5)	1350 (7)	90 (<1)	150 (24)	-
CT evidence of infarction	6331 (34)	-	15140 (75)	2397 (12)	103 (17)	-

Table 6-4 Early and long term outcome events in aspirin trials.

	IST-1 (N = 18372)		CAST (N = 20172)		MAST-I (N = 622)		Total (N = 39166)	
	No.	%	No.	%	No.	%	No.	%
Early outcomes, 14 days or less								
Deep venous thrombosis (DVT)	21	0.11	-	-	-	-	21	0.05
Pulmonary embolism (PE)	122	0.66	19	0.09	-	-	141	0.36
Ischaemic stroke/ cerebral infarction	632	3.44	319	1.58	5	0.80	956	2.44
Myocardial infarction (MI)	357	1.94	49	0.24	5	0.80	411	1.05
All thrombotic events	**1118**	**6.09**	**380**	**1.88**	**10**	**1.62**	**1508**	**3.85**
Significant intracranial haemorrhage	119	0.65	75	0.37	32	5.14	226	0.58
Major extra-cranial haemorrhage	150	0.82	90	0.45	8	1.29	248	0.63
Haemorrhagic transformation	49	0.27	50	0.25	-	-	99	0.25
All haemorrhagic events	**265**	**1.44**	**204**	**1.01**	**36**	**5.79**	**505**	**1.29**
Long term outcome, six months or less								
Common functional score (1-4)								
(1) Recovered	3142	17.10	7415	36.76	167	26.85	10724	27.38
(2) Not recovered	3680	20.03	6498	32.21	59	9.49	10237	26.14
(3) Dependent	7460	41.61	5445	26.99	209	33.60	13114	33.48
(4) Dead	3953	22.52	765	3.79	187	30.06	4905	12.52
Dead or dependent	**11413**	**62.12**	**6210**	**30.79**	**396**	**63.67**	**18019**	**46.01**
Missing	137	0.75	49	0.24	-	-	186	0.01

Note: Twenty one individuals experienced both a haemorrhagic and a thrombotic event in IST-1.

6.2.4.3 Missing data within each of the trials

Baseline characteristics for the three datasets are summarised in Table 6-3 alongside the proportion and frequency of missing data. The MAST-I data were entirely complete for these measures with the exception of systolic blood pressure with three missing readings. Atrial fibrillation was not recorded for any patient during the initial pilot recruitment phase of IST-1 (January 1991 to February 1993). These observations can therefore be assumed to be Missing At Random (MAR). The IST-1 had 12291 (67%) patients with completely observed data, 2825 (15%) with one missing value and 1699 (9%) with two. Fewer had multiple missing values with one patient having a maximum of ten missing values. The CAST data had a total of 15657 (78%) patients with completely observed data, 2406 (12%) with one missing and 910 (5%) with two. As with the IST-1 data, fewer had multiple missing values with one patient having a maximum of eleven missing values.

Joint missingness was explored using cluster plots (Figure 6-1 and Figure 6-2) for the IST-1 and CAST trial datasets (NB such a plot could not be produced for MAST-I since there were no jointly missing values). In both the IST-1 and the CAST data, the majority of jointly missing values were associated amongst the stroke deficits.

Imputation was undertaken for each of the datasets. The adopted imputation model included each of the variables listed in the cluster plots (Figure 6-1 and Figure 6-2). Indicator variables for missingness amongst each variable can be created to explore the extent to which missingness can be explained by the observed values. For IST-1 this ranged from as low as 3% to as high as 90% and for CAST as low as 3% to as high as 82%. Missingness may be at least partly Missing At Random (MAR) and therefore a multiple imputation approach could reduce the risk of bias. Missing data were imputed multiple times generating twenty complete datasets per trial.

It is noted that 186 (0.01%) of the 39166 patients have missing long term outcomes. Imputing missing outcomes contributes random noise to the results (Von Hippel, 2007); however, given how few had missing outcomes, the expected impact of this added noise is negligible.

Figure 6-1 Combinations of missing values in IST-1; a hierarchical cluster analysis of combined missingness.

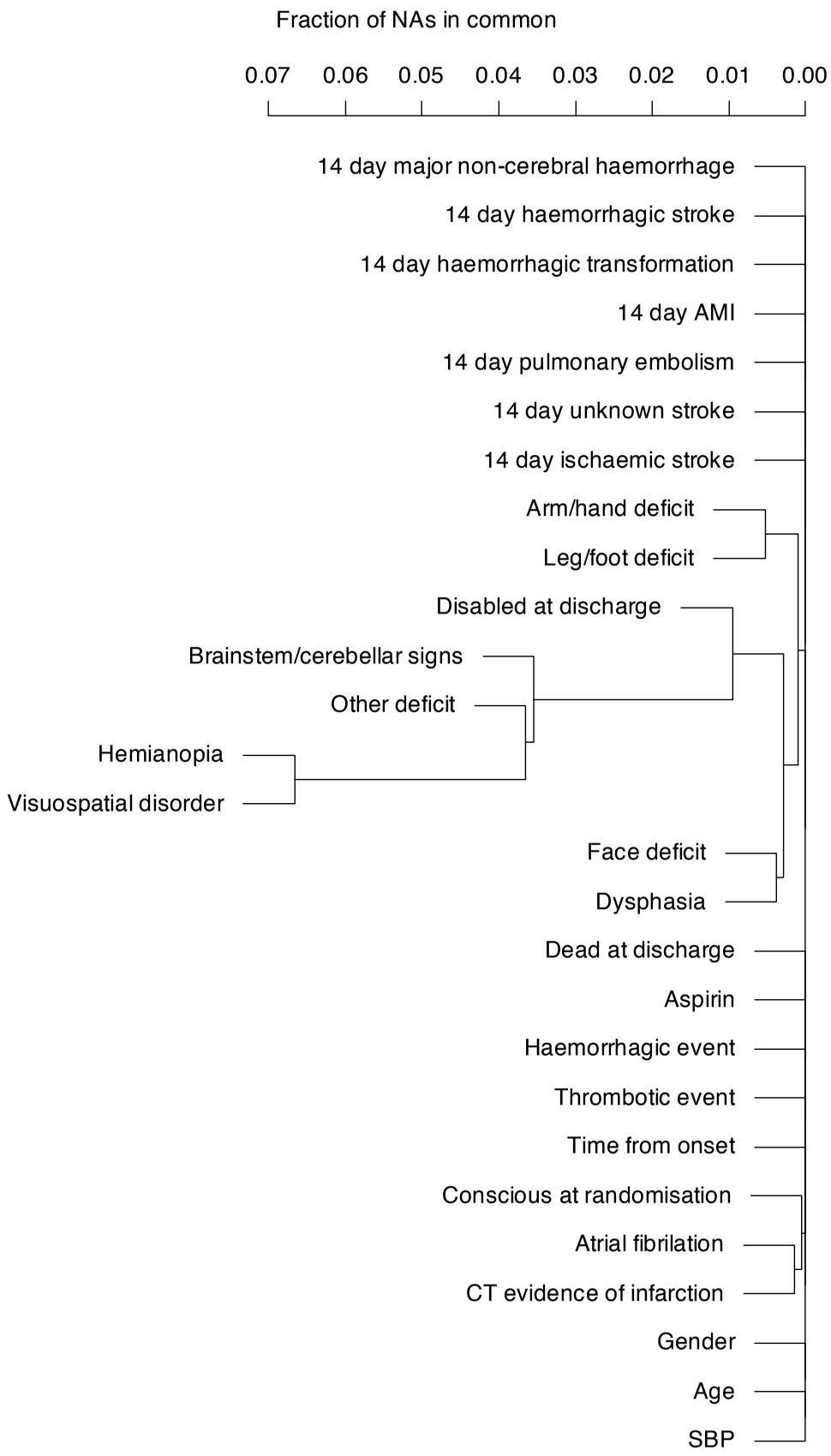


Figure 6-2 Combinations of missing values in CAST; a hierarchical cluster analysis of combined missingness.

6.2.5 Pooled aggregate results from aspirin trials

The IST-1 and CAST trials were run concurrently. The intention was that both analyses would be pooled to show the effect of aspirin within a much larger sample of ischaemic stroke patients than had ever been studied before (Chen et al., 2000). Following the Chen *et al* paper in 2000, a Cochrane analysis was conducted in 2008 using all three trials (CAST, IST-1 and MAST) (Sandercock et al., 2008). The figure below describes how the three datasets were combined with respect to: (i) the Cochrane analysis (dotted lines); and (ii) the analyses undertaken in this chapter (solid lines).

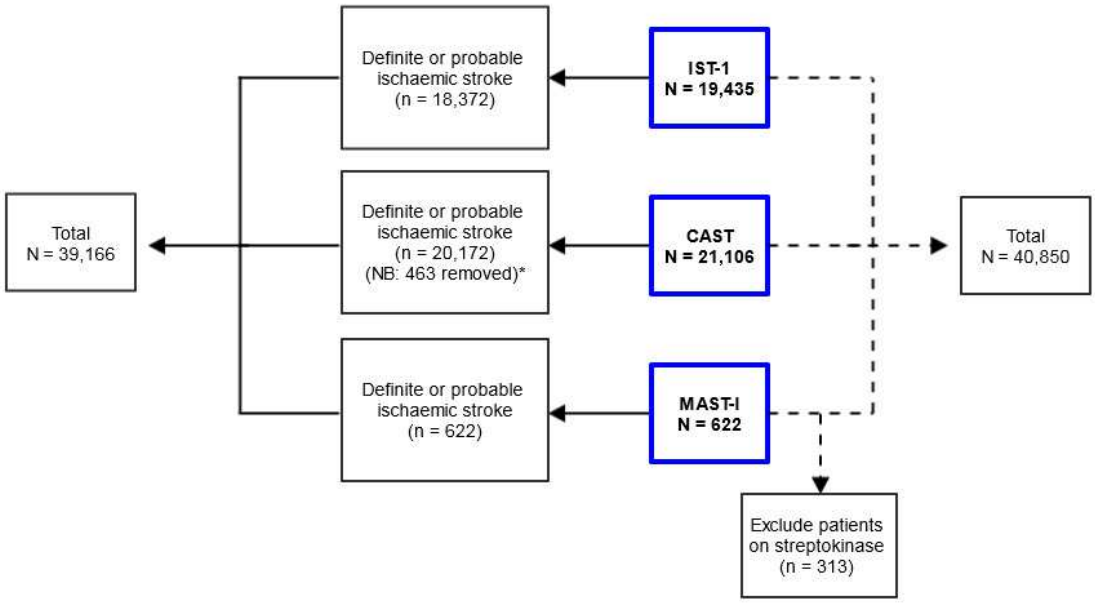
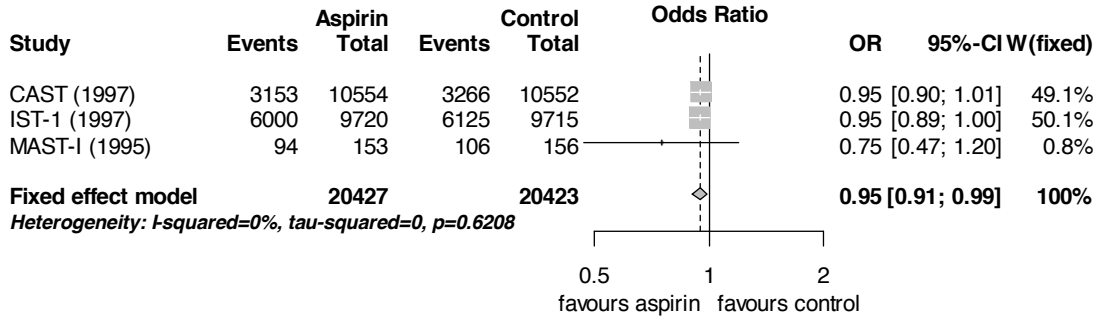


Figure 6-3 Flow-diagram of included aspirin trial IPD. Solid lines indicate how the datasets are handled in this chapter whilst dotted lines indicate the original Cochrane analysis. Note '*' the obtained patient data for CAST did not match publication

The Cochrane review considered all patients in IST-1 treated with aspirin since there was no indication that heparin and aspirin interacted with one another to produce a difference in outcome, harm or benefit. In the case of MAST-I there was an interaction between treatments. It was noted above that the trial was stopped early due to an excess number of early deaths (see section 6.2.3). Sandercock *et al*. therefore exclude those patients randomised to receive streptokinase.

A: Cochrane meta-analysis



B: Analysis using obtained IPD

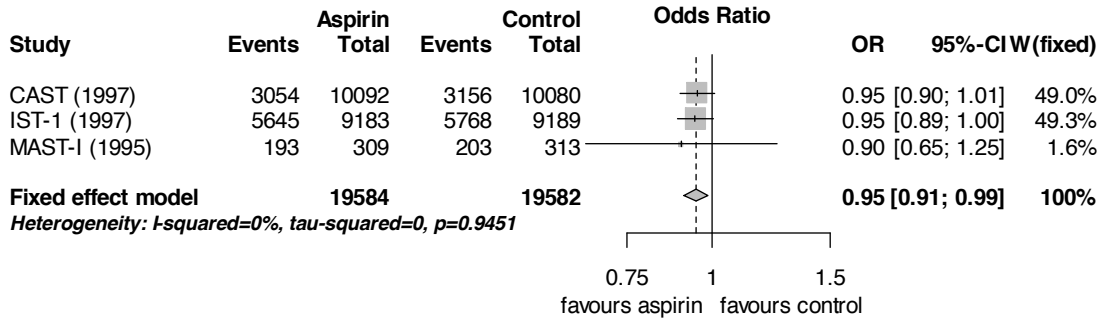


Figure 6-4 Meta-analysis of three RCTs assessing aspirin versus control: (A) as per Cochrane review data; and (B) using available IPD for the analysis in this chapter (note 137 and 49 patients missing in IST-1 and CAST respectively)

This analysis is reproduced in Figure 6-4 (A) above and contrast with the data used specifically in this chapter Figure 6-4 (B). For this chapter only confirmed ischaemic or unknown strokes at baseline are included and MAST-I is used in its entirety. The addition of those on streptokinase resulted in a slightly smaller effect from aspirin although in neither setting was this shown to be significantly different from no effect. The pooled result is identical and the reproducibility of the data is verified.

6.3 Developing and evaluating new models

Models for the prediction of recurrent stroke or myocardial infarction were identified in Chapter 3. None of which, however, could be applied to the IST-1, CAST or MAST-I datasets since the required baseline predictors of these models were not all recorded by these trials. Therefore, new models had to be developed. In order to both *develop* and *evaluate* new clinical prediction models a non-random split of the IST-1 trial data was used. This was done by country of recruitment and constitutes a geographical evaluation of performance (Toll et al., 2008). A simple random-split ensures similarly distributed characteristics in each sample and is overly optimistic. Patients recruited to UK and Italian hospitals made up the *training set* and were used for model development whilst the remaining 34 countries formed the *test set* and were used for model evaluation. Development was undertaken in the training set.

6.3.1 Predicting 14 day thrombosis and haemorrhage

Two multivariable binary logistic regression (LR) models for the separate prediction of: 14 day thrombosis on control; and 14 day haemorrhage on aspirin were developed. Linearity and additivity assumptions were assessed using restricted cubic splines and two-way interactions respectively. A single imputed dataset was generated using the imputation model previously discussed. The model development subset (i.e., the *training set*) was specified within this single imputation and the model assumptions were explored. Clinical plausibility was given prominence over statistical significance. A high level of improvement had to be demonstrated before the added complexity of either an interaction or a non-linear association would be included in the final models. To meet inclusion an increase of 10 or more Akaike's Information Criterion (AIC) units on the Chi-squared scale was set (Steyerberg, 2009). The changes in AIC units are presented in Table 6-5. There were no substantial improvements in either of the early event models when including non-linear effects nor when considering the inclusion of interaction terms. Increasing age and the presence of a visible infarct were the only statistically significant prognostic factors; both suggested an increase in the log-odds of thrombosis (Table 6-6). For most of the predictors the association was in the same direction across each of the events. Sensitivity analyses are provided in section 6.6 Appendix A on page 177.

Table 6-5 Testing model assumptions for early event models in a single imputed dataset.

Test performed	Age (years)	SBP (mmHg)	OTT (hours)
Fourteen day thrombosis model			
Additivity			
Gender	-1.84 (0.6934)	-1.81 (0.6664)	0.86 (0.0908)
Visible infarct on CT	-1.28 (0.3968)	-0.09 (0.1668)	-2.00 (0.9607)
Atrial fibrillation	-1.21 (0.3753)	-1.94 (0.8106)	-2.00 (0.9842)
Conscious (Drowsy/Coma vs. Alert)	-0.38 (0.2033)	-0.25 (0.1862)	-1.95 (0.8211)
Motor deficit	-1.73 (0.6044)	-1.99 (0.9042)	-2.00 (0.9672)
Presence of dysphasia	-1.28 (0.3957)	0.48 (0.1154)	-1.94 (0.8106)
Presence of hemianopia	-0.68 (0.2505)	0.85 (0.0911)	-1.35 (0.4211)
Presence of visuospatial disorder	-0.36 (0.2006)	-2.00 (0.9992)	-1.19 (0.3694)
Presence of brainstem/ cerebellar deficit	1.41 (0.0647)	-1.96 (0.8423)	-1.82 (0.6746)
Linearity			
Restricted cubic spline			
Three knots	1.41 (0.0648)	0.15 (0.1426)	-1.95 (0.8258)
Four knots	-0.38 (0.1633)	2.37 (0.0413)	-3.95 (0.9766)
Five knots	-2.36 (0.3031)	0.87 (0.0763)	-1.06 (0.1761)
Square term	1.38 (0.0662)	0.89 (0.0891)	-1.99 (0.9251)
Fourteen day haemorrhage model			
Additivity			
Gender	-1.50 (0.4805)	-0.66 (0.2466)	-1.18 (0.3651)
Visible infarct on CT	-1.77 (0.6315)	-1.29 (0.4005)	-1.97 (0.8565)
Atrial fibrillation	-0.88 (0.2890)	-1.97 (0.8634)	-1.91 (0.7613)
Conscious (Drowsy/Coma vs. Alert)	-0.01 (0.1588)	-1.63 (0.5404)	-1.97 (0.8682)
Motor deficit	0.45 (0.1177)	-2.00 (0.9525)	-2.00 (0.9819)
Presence of dysphasia	-1.39 (0.4344)	-1.99 (0.9296)	-2.00 (0.9742)
Presence of hemianopia	-1.97 (0.8733)	1.41 (0.0648)	-1.93 (0.7882)
Presence of visuospatial disorder	-1.99 (0.9065)	0.99 (0.0839)	-1.86 (0.7036)
Presence of brainstem/ cerebellar deficit	1.57 (0.0589)	-1.44 (0.4531)	-2.00 (0.9685)
Linearity			
Restricted cubic spline			
Three knots	-0.48 (0.2171)	-1.84 (0.6901)	-1.75 (0.6164)
Four knots	-2.26 (0.4185)	-3.52 (0.7849)	-0.47 (0.1715)
Five knots	-2.72 (0.3507)	-5.06 (0.8162)	-1.57 (0.2185)
Square term	0.40 (0.1213)	-1.91 (0.7618)	-1.49 (0.4770)

NOTE: The difference in AIC units contrasting a complex fit to a simple fit is provided alongside the P-value for the associated LR test. Positive AIC on the chi-squared scale represent an improved fit

Table 6-6 Multivariable prediction models for 14 day events in the development split with imputed IST-1 data (over 20 imputed sets)

	Thrombotic model (291/4504)			Haemorrhagic model (74/4511)		
Variable	**Coefficient (SE)**	**OR (95% CI)**	**P-value**	**Coefficient (SE)**	**OR (95% CI)**	**P-value**
Intercept	-4.339 (0.659)	-	<0.0001	-5.8526 (1.286)	-	<0.0001
Age (per decade)	0.193 (0.064)	1.21 (1.07 to 1.38)	0.0027	0.215 (0.125)	1.24 (0.97 to 1.58)	0.0860
SBP (per 10 mmHg)	-0.024 (0.023)	0.98 (0.93 to 1.02)	0.3040	-0.015 (0.045)	0.98 (0.90 to 1.08)	0.7362
Delay from randomisation (hours)	-0.009 (0.005)	0.99 (0.98 to 1.00)	0.0641	-0.005 (0.010)	1.00 (0.98 to 1.01)	0.6422
Sex (Male)	0.201 (0.127)	1.22 (0.95 to 1.57)	0.1125	0.379 (0.246)	1.46 (0.90 to 2.37)	0.1239
Visible infarct on CT	0.419 (0.135)	1.52 (1.17 to 1.98)	0.0019	0.110 (0.271)	1.12 (0.66 to 1.90)	0.6841
Conscious (Drowsy/Coma vs. Alert)	-0.028 (0.162)	0.97 (0.71 to 1.34)	0.8639	0.001 (0.315)	1.00 (0.54 to 1.86)	0.9986
Atrial fibrillation	0.168 (0.153)	1.18 (0.88 to 1.60)	0.2743	0.216 (0.314)	1.24 (0.67 to 2.30)	0.4924
At least one or more of arm/leg deficits	0.431 (0.226)	1.54 (0.99 to 2.39)	0.0561	0.309 (0.450)	1.36 (0.56 to 3.28)	0.4937
Presence of dysphasia	0.026 (0.131)	1.03 (0.79 to 1.33)	0.8441	-0.361 (0.255)	0.70 (0.42 to 1.15)	0.1567
Presence of hemianopia	0.159 (0.187)	1.17 (0.81 to 1.69)	0.3949	0.334 (0.366)	1.40 (0.68 to 2.86)	0.3607
Presence of visuospatial disorder	0.103 (0.188)	1.11 (0.77 to 1.60)	0.5831	0.036 (0.351)	1.04 (0.52 to 2.06)	0.9193
Presence of brainstem/ cerebellar deficit	0.341 (0.202)	1.41 (0.95 to 2.09)	0.0917	-0.338 (0.484)	0.71 (0.28 to 1.84)	0.4844

6.3.2 Evaluating the performance of the early event models

The two multivariable logistic regression models developed for the prediction of 14 day thrombosis and 14 day haemorrhage (Table 6-6) were internally evaluated using the *test set* (i.e., non UK/Italian recruited hospital patients). Model performance was assessed using: (i) *Nagelkerke's* R^2, an overall measure of model performance; (ii) the AUROCC to quantify model discrimination where perfect discrimination has an AUROCC of 1 and no discrimination an AUROCC of 0.5; and (iii) calibration, split into the calibration intercept and slope, which should ideally be as close to 0 and 1 respectively as possible. These measures were calculated on the twenty imputed *test sets* and combined using Rubin's rules (Table 6-7). Discrimination between those patients with and without thrombosis by 14 days (AUROCC of 0.56) and those with and without haemorrhage by 14 days (AUROCC of 0.57) was poor. Calibration was moderate, although for both outcomes there was the suggestion that the outcome-predictor associations in development were optimistically large (slope<1) as well as a slight systematic component indicating a consistent under prediction of risk from early events (intercept>0).

Table 6-7 Performance of multivariable prediction models for 14 day events in the evaluation split with imputed test data. Note 95%CI provided unless otherwise stated

Performance Measure	Thrombosis (337/4685)	Haemorrhage (87/4672)
Median % R^2 (IQR)	0.58 (0.55 to 0.59)	0.57 (0.39 to 0.67)
AUROCC	0.56 (0.53 to 0.59)	0.57 (0.52 to 0.64)
Calibration		
Intercept	0.14 (0.09 to 0.20)	0.20 (0.09 to 0.31)
Slope	0.46 (0.33 to 0.60)	0.52 (0.26 to 0.78)

Note: Inter-quartile range (IQR) is based on 20 imputations

6.3.3 Predicting six month death or dependency

The same methodological approach used for the development of prediction models for early events was used to develop a model to predict six month death or dependency on control. Two non-linear effects were identified as important for age and systolic blood pressure. This departure from linearity was identified first by fitting restricted cubic splines were it was observed that a substantial improvement in model fit could be made through this flexible fit (Table 6-8). However it was also noted that the same improvement could be made by fitting a simple quadratic function (Figure 6-5).

The presence of arm/leg weakness, dysphasia, hemianopia, visuospatial disorder and unconscious or drowsy at randomisation were associated with a statistically significant increase in the log-odds of death or dependency (Table 6-9).

Table 6-8 Testing model assumptions for six month death or dependency model in a single imputed dataset.

Test performed	Age (years)	Systolic BP (mmHg)	OTT (hours)
Additivity			
Gender	1.95 (0.8261)	-1.74 (0.6082)	-1.23 (0.3793)
Visible infarct on CT	-1.97 (0.8711)	-2.00 (0.9903)	7.20 (0.0024)
Atrial fibrillation	5.30 (0.0069)	-1.53 (0.4923)	-1.78 (0.6382)
Conscious (Drowsy/Coma vs. Alert)	1.06 (0.0802)	-1.41 (0.4429)	-1.96 (0.8323)
Motor deficit	-1.89 (0.7364)	-0.83 (0.2794)	-2.00 (0.9570)
Presence of dysphasia	-0.49 (0.2193)	-1.95 (0.8294)	-1.69 (0.5780)
Presence of hemianopia	3.52 (0.0188)	1.93 (0.0475)	-1.34 (0.4151)
Presence of visuospatial disorder	-0.77 (0.2680)	-2.00 (0.9633)	-1.58 (0.5151)
Presence of brainstem/ cerebellar deficit	-1.92 (0.7717)	-0.66 (0.2464)	-1.53 (0.4937)
Linearity			
Restricted cubic spline			
Three knots	**41.18 (<0.0001)**	**10.47 (0.0004)**	0.85 (0.0916)
Four knots	**42.07 (<0.0001)**	8.61 (0.0018)	-0.80 (0.2018)
Five knots	**43.10 (<0.0001)**	6.63 (0.0055)	0.10 (0.1069)
Square term	**40.35 (<0.0001)**	**10.63 (0.0004)**	1.50 (0.0615)

NOTE: The difference in AIC units contrasting a complex fit to a simple fit is provided alongside the P-value for the associated LR test. Positive AIC on the chi-squared scale represent an improved fit

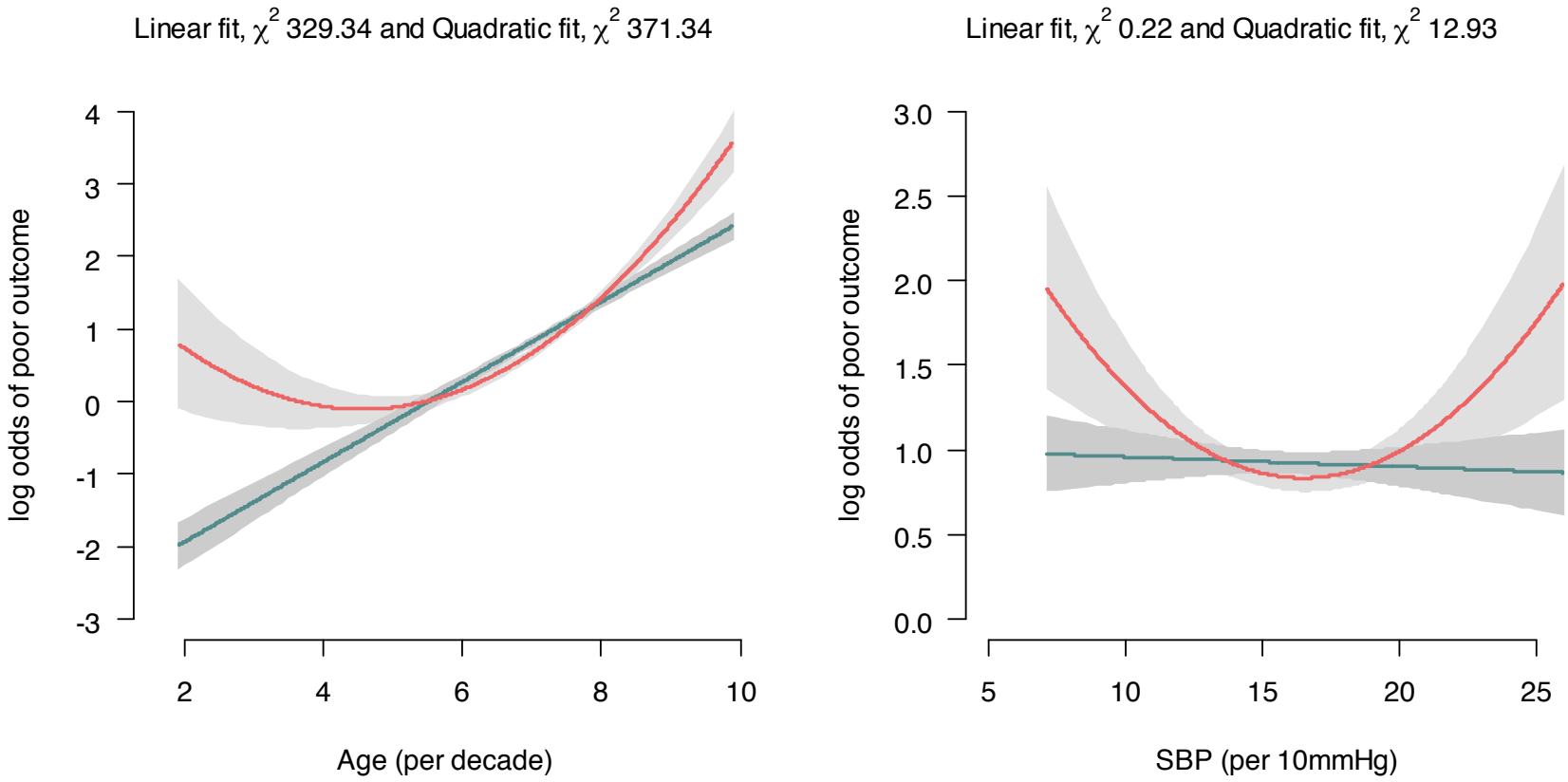


Figure 6-5 Transformations of patient age (per decade) and systolic blood pressure (per 10mmHg) in univariable analysis comparing a simple linear fit to a quadratic function

Table 6-9 Multivariable prediction models for six month death or dependency in the development split with imputed IST-1 data (over 20 imputed sets)

	Development split (3225/4504)		
Variable	**Coefficient (SE)**	**OR (95% CI)**	**P-value**
Intercept	3.098 (1.384)		0.0252
Age (per 10 years)			
Age	-0.801 (0.293)	0.44 (0.25 to 0.79)	0.0063
Age^2	0.095 (0.022)	1.10 (1.05 to 1.15)	<0.0001
SBP (per 10 mmHg)			
SBP	-0.359 (0.136)	0.69 (0.53 to 0.91)	0.0083
SBP^2	0.011 (0.004)	1.01 (1.00 to 1.02)	0.0058
Delay from randomisation (hours)	0.004 (0.003)	1.00 (1.00 to 1.01)	0.2001
Sex (Male)	-0.329 (0.078)	0.72 (0.62 to 0.84)	<0.0001
Visible infarct on CT	0.051 (0.088)	1.05 (0.89 to 1.25)	0.5638
Conscious (Drowsy/Coma vs. Alert)	1.495 (0.147)	4.48 (3.36 to 5.98)	<0.0001
Presence of atrial fibrillation	0.109 (0.113)	1.11 (0.89 to 1.38)	0.3342
At least one or more of arm/leg deficits	0.937 (0.109)	2.58 (2.08 to 3.19)	<0.0001
Presence of dysphasia	0.365 (0.080)	1.44 (1.24 to 1.69)	<0.0001
Presence of hemianopia	0.732 (0.132)	2.06 (1.59 to 2.67)	<0.0001
Presence of visuospatial disorder	0.618 (0.130)	1.83 (1.41 to 2.38)	<0.0001
Presence of brainstem/ cerebellar deficit	-0.176 (0.122)	0.83 (0.66 to 1.06)	0.1497

6.3.4 Evaluating the performance of the death or dependency model

The six month death or dependency model discriminated well between those with poor functional outcome (dead or dependent) and those without (Table 6-10) within the *test dataset*, the CAST and the MAST-I data (AUROCC values ranging from 0.77 to 0.71). Calibration was moderate and reflected some key differences between incidence of outcome and trial design. The observed calibration slope in internal evaluation and in the CAST evaluation (<1) suggests optimistically large predictor-outcome associations in model development indicative of overfitting, whilst the negative intercepts points toward an over estimation in the predicted risks. This can be explained by a shorter follow-up period in CAST (28 days compared to 6 months in IST-1) and a greater incidence of death or dependence amongst the UK and Italian recruitment centres compared to the *test dataset* (72% versus 55%). The similar incidence rate in MAST-I would suggest good calibration-in-the large which indeed was near equal to zero. The slope suggests smaller predictor-outcome associations

The model had similar discrimination and calibration in the CAST (AUROCC 0.71, slope 0.92 and intercept -1.31) and MAST-I trials (AUROCC 0.76, slope 1.39 and intercept 0.02). The calibration slope in MAST-I reflects the earlier follow-up (28 days) when contrast to IST-1 and MAST-I.

Table 6-10 Performance of six month death or dependence model in test set and CAST and MAST-I (over 20 imputed sets)

	Data set		
Performance Measure	Test data: 2587/4685	CAST: 3164/10080	MAST: 203/313
% R^2 (IQR)	28.58 (28.46 28.82)	17.14 (17.09 to 17.22)	24.93 (24.90 to 24.98)
AUROCC	0.77 (0.76 to 0.78)	0.71 (0.70 to 0.72)	0.76 (0.70 to 0.81)
Calibration			
Intercept	-0.75 (-0.78 to -0.71)	-1.31 (-1.34 to -1.29)	0.02 (-0.10 to 0.14)
Slope	0.90 (0.87 to 0.93)	0.92 (0.89 to 0.94)	1.39 (1.19 to 1.59)

Note: Inter-quartile range (IQR) is from 20 imputations

6.4 Heterogeneity in treatment effect

In this section, the predicted risk of early and late events is calculated for each patient in each of the three trials. Differences in incidence will be corrected using a simple updating procedure. Differences in the effect of aspirin on poor functional outcome will be explored on both the absolute and the relative scale. A single imputed set for missing values for each trial were used for practicality and interpretability; although a more rigorous approach would use each of the imputed datasets and pool the results.

6.4.1 Recalibration and updating of prediction models

Model calibration describes how accurately a given model estimates risk. Ideally predictions should closely match that which is observed. The three binary logistic regression models developed using the IST-1 dataset should, in principle, be suitable for use in new patients similar to those in the UK and Italian recruitment hospitals of the development set, i.e., similar baseline characteristics and outcome rates. In such instances these models should produce accurate predicted risks and may be described as *well calibrated*. However, it is evident from Table 6-4 that, for example, the incidence of thrombotic events differ from trial to trial (6.1%, 1.9% and 1.6% in IST-1, CAST and MAST-I respectively). Some adjustment must therefore be made to enable sensible predictions for each patient within each trial. If not, all patients will have predicted risk estimates that match those seen in the development data which are known to not be representative of the other two trials. A simple re-calibration of the three binary logistic regressions intercepts was carried out adjusting for the differences in incidence between the development data and the new data (Janssen et al., 2009). This form of adjustment relates only to the estimated risk and does not affect the ordering of the patients. Thus, discrimination, a rank based measure, is unaffected. A correction factor is calculated from the incidence of the outcome in the new data (n/N) and the mean predicted risk estimated by the model ($\bar{\hat{p}}$).

$$cf = \log\left(\frac{n/N}{1-(n/N)} \Big/ \frac{\bar{\hat{p}}}{1-\bar{\hat{p}}} \right) \qquad \textbf{(6.1)}$$

This correction factor (*cf*) can be interpreted as the log odds ratio of incidence and estimated risk. Adding this *cf* to the original model constitutes as updating the old model to the new local setting (Table 6-11).

Table 6-11 Models recalibrated for IST-1, CAST and MAST-I. Note intervals provided are 95%CIs

Measure	Thrombosis	Haemorrhage	Dead/dependent
IST-1			
Correction factor (*cf*)	0.078	0.102	-0.313
events/Total	628/9189	161/9183	5812/9189
Calibration intercept	0.00 (-0.08 to 0.08)	0.00 (-0.15 to 0.16)	-0.08 (-0.13 to -0.04)
Calibration slope	0.68 (0.47 to 0.89)	0.71 (0.32 to 1.09)	0.94 (0.89 to 0.99)
CAST			
Correction factor (*cf*)	-1.105	-0.032	-1.135
events/Total	206/10080	123/10092	3164/10080
Calibration intercept	-0.01 (-0.14 to 0.13)	0.00 (-0.18 to 0.18)	-0.18 (-0.23 to -0.14)
Calibration slope	0.81 (0.41 to 1.21)	0.42 (-0.08 to 0.92)	0.92 (0.87 to 0.97)
MAST-I			
Correction factor (*cf*)	-1.695	1.446	0.047
events/Total	4/313	22/309	203/313
Calibration intercept	-0.01 (-0.99 to 0.98)	0.01 (-0.43 to 0.44)	0.02 (-0.27 to 0.22)
Calibration slope	0.37 (-2.73 to 3.47)	-0.88 (-2.08 to 0.33)	1.39 (1.00 to 1.79)

Note: correction factors are based on arm specific incidences

6.4.2 Predicted risk of early events across the three trials

Each patient's risk from an early event was predicted using the two logistic regression models updated for each trial using the adjustment method given by equation (6.1). Figure 6-6 given below adequately captures all elements of this problem. For each trial patients predicted risk of haemorrhage and thrombosis were plotted. Due to the size of the IST-1 and CAST trials a simple random sample of 600 patients (drawn from each of the trial totals) was used so as to aid visual interpretation. The plots that appear in the left hand column are of those patients on aspirin and the right hand column of those on control. The *x* and *y*-axes have different scales for each trial, reflecting the differences in early event rates. For example, in CAST the rate of thrombotic events was much lower than that seen in IST-1 explaining why the data scatter lies closer to the line of equality (dashed diagonal line) whereas for MAST-I there was a greater incidence of haemorrhage, thus the scatter falls below the line of equality. The grey vertical and horizontal lines highlight quarters of predicted haemorrhagic and thrombotic risk respectively. These were defined using the IST-1 trial with cut-points made at the 25th, the 50th and the 75th percentiles. This categorisation was then applied in turn to each of the trials thus defining sixteen distinct strata of predicted risk. The sixteen strata were not filled equally by each of the trials, for instance, in MAST-I all patients fell within the strata comprising the upper haemorrhagic quarter and lower thrombotic quarter. Patients were plotted as dark grey if dead or dependent by the end of follow-up. The difference in follow-up was therefore captured in the CAST plots as incidence of death or dependency was about half of that seen in IST-1 and MAST-I. This figure is of vital importance. It enables a visual representation of the problem in its entirety. The distribution of those dead or dependent across predicted risk is largely the same in each trial and each arm. Trials contribute evidence differently according to their predicted risks of early events, populating the sixteen strata to varying degrees. A correlation between predicted risk of thrombosis and haemorrhage can be seen in each of the plots highlighting that the identification of thrombotic events from haemorrhagic events is poor. Finally, there are no identifiable patterns of death or dependence across the sixteen defined risk strata nor are there any such patterns evident under some other possible categorisation.

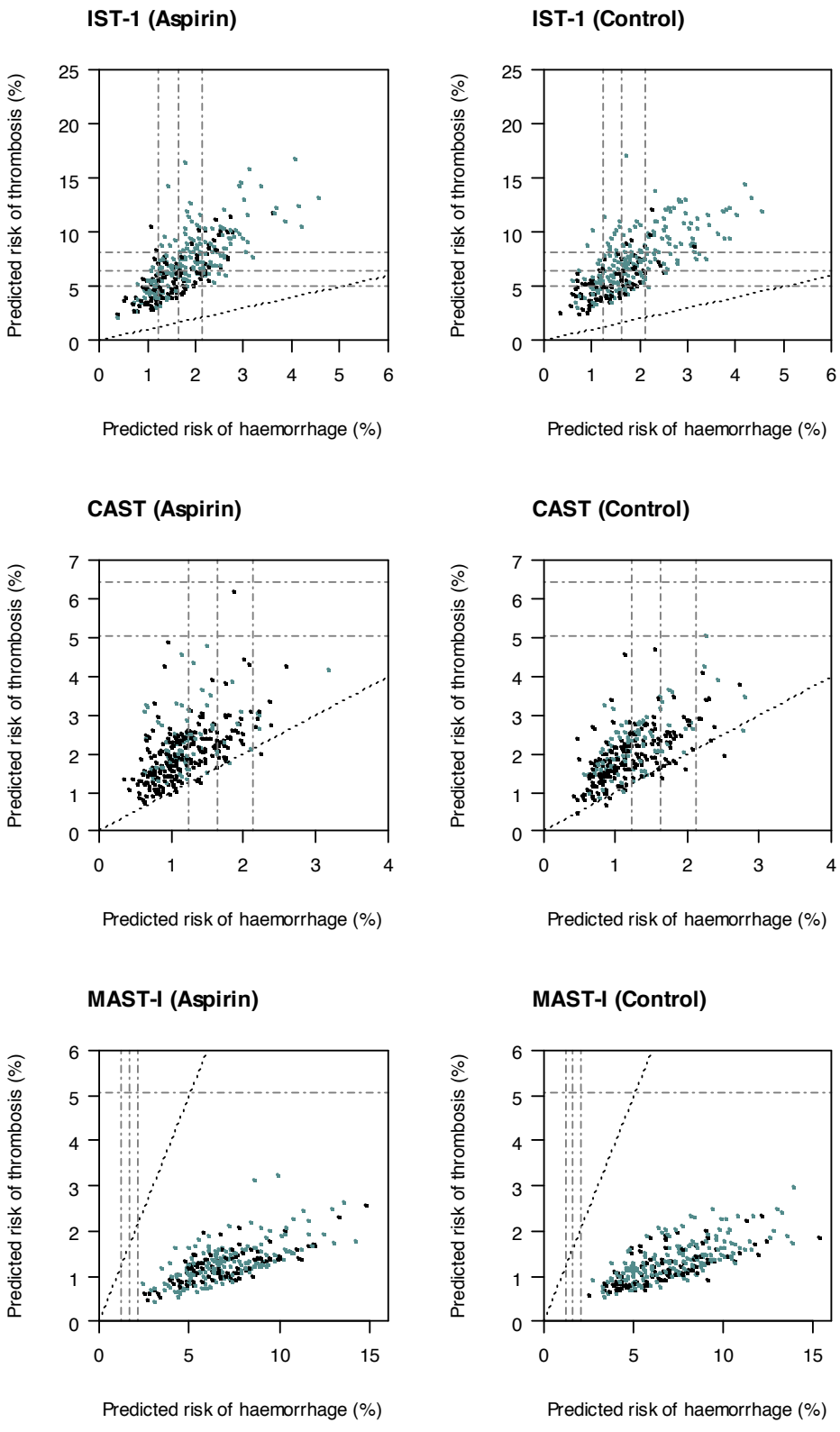


Figure 6-6 Predicted risk of thrombosis vs. predicted risk of haemorrhage. Horizontal and vertical grey lines indicate quarters of risk. Grey points indicate patient dead or dependent and black alive and independent. Points are a random sample of 600 patients from each trial

6.4.2.1 Meta-analysis of treatment effect within predicted risk strata

The meta-analysis approach used pools together all of the information displayed in Figure 6-6. The absolute risk reduction of poor functional outcome within each of the sixteen risk groups for each trial was estimated. Brown and Prescott state that it is inadvisable to model random effects when the number of centres (trials) is anything less than five (Brown and Prescott, 2006). Each estimate was then pooled across the three trials using a fixed effects Mantel-Haenszel meta-analysis. Out of all sixteen of these risk strata there was only one occasion (see MAST-I scatter plots in Figure 6-6) where all three studies provided an observed ARR. It is again emphasised that for some of the risk strata estimates of absolute risk reduction in poor functional outcome were obtained from the IST-1 data only (Figure 6-7).

Inspection of the 16 groups (Figure 6-7 above) suggests that there is no evidence that those with a high haemorrhagic risk and a low thrombotic risk experience any overall harm from treatment with aspirin, i.e., an excess of death or dependency on aspirin, nor was there evidence that those with a low haemorrhagic risk and high thrombotic risk benefited any more than the overall estimate of a reduction in absolute risk of poor functional outcome by 1%. In no risk group of haemorrhage or thrombotic events was the absolute reduction in poor functional outcome at final follow up statistically different from the overall estimate of 1%. Only one out of the sixteen pooled strata specific ARR estimates was based on IST-1, CAST and MAST-I data. Nine were based on both IST-1 and CAST, and six on IST-1 data alone.

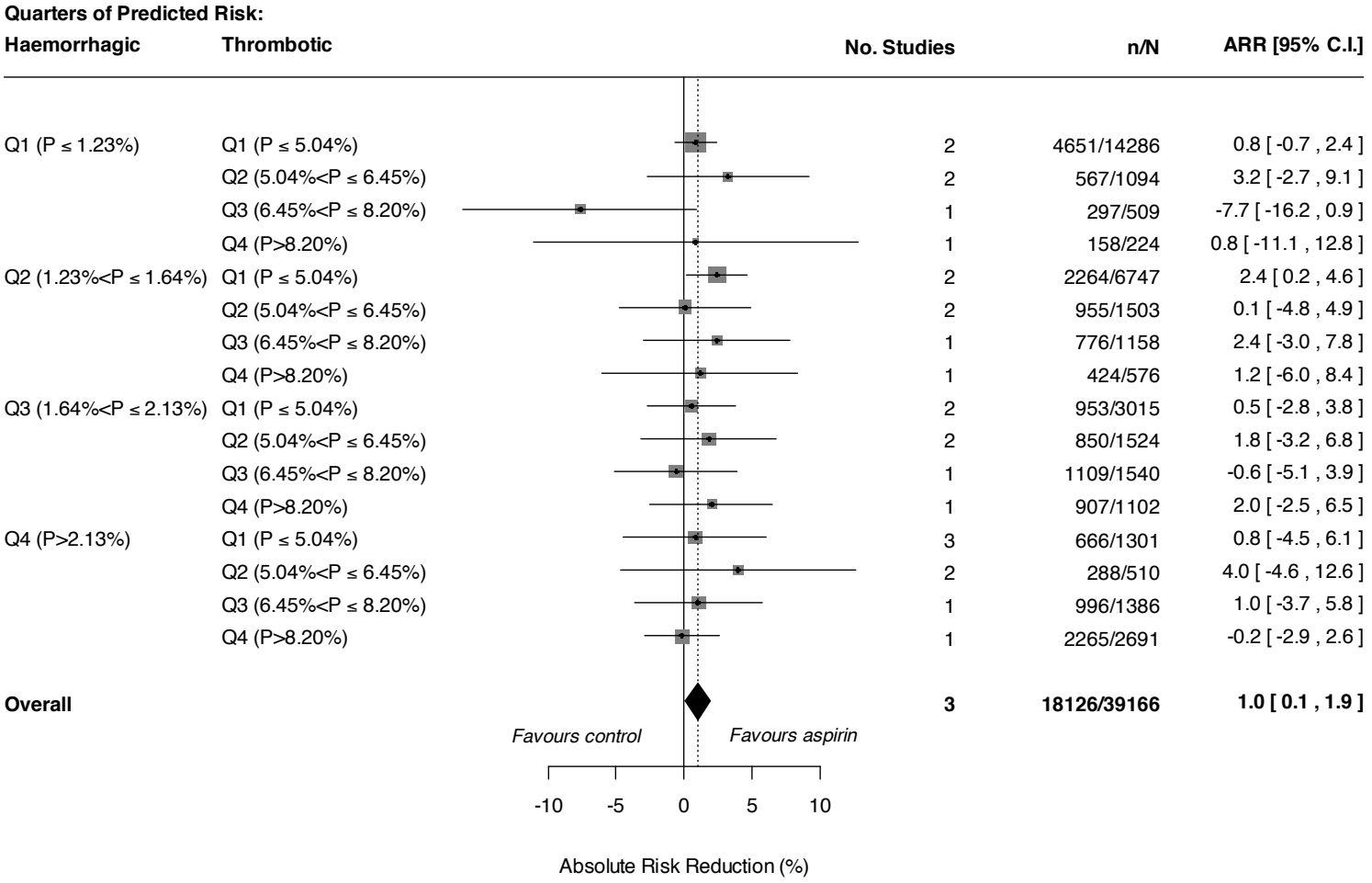


Figure 6-7 Pooled estimates of ARR in poor outcome. Q1 to Q4 denote quarters of risk. Square sizes are proportional to the strata specific denominator

6.4.2.2 Relative effect of aspirin by predicted risk of early events

Predicted probabilities of patient risk from 14 day thrombosis and 14 day haemorrhage were calculated for all 39166 patients. These were fitted as continuous predictors within a binary logistic regression for the outcome 'death or dependency' at last follow-up and also within a proportional odds logistic regression modelling the full ordinal outcome scale of death or disability (defined in Table 6-2). For each of these models the predicted risks were introduced on the original predicted log odds scale. To enable a sensible biological interpretation of the interaction between risk of early events and treatment with aspirin, each of the predicted early outcome terms were centred on their respective means (Schielzeth, 2010). The table below describes the contribution of each covariate to the full model fit. An adjustment was made for *trial* which made the largest contribution to the model fit (binary and ordinal). This is likely due to the differences in follow-up between CAST and IST-1 Table 6-4. Whilst the predicted risk of thrombosis and haemorrhage provided sizable contributions to each of the respective model fits, there was no evidence to support an interaction between these covariates and aspirin so far as separate two way interactions or three way interactions.

For the binary LR fit aspirin appeared to reduce the odds of poor outcome although it was not statistically significant (OR 0.96 95% CI 0.92 to 1.01). The more sensitive proportional odds LR did demonstrate a statistically significant shift in the ordinal scale favouring better functional outcomes (OR 0.95 95% CI 0.91 to 0.99).

Note that the exclusion of those patients randomised to streptokinase or high dose heparin made no major difference to these results (see section 6.6 Appendix A on page 177).

Table 6-12 Wald statistics for a binary LR (dead or dependent) and a proportional odds LR (across the ordinal functional outcome) modelling predicted risks from early events

	Binary LR			Proportional Odds LR		
Factor	**χ^2**	***d.f.***	**P-value**	**χ^2**	***d.f.***	**P-value**
Predicted haemorrhagic risk (X_H)	33.10	4	<0.0001	63.95	4	<0.0001
All interactions	11.30	3	0.0102	52.44	3	<0.0001
Predicted thrombotic risk (X_T)	1436.09	4	<0.0001	1910.44	4	<0.0001
All interactions	13.47	3	0.0037	54.68	3	<0.0001
Aspirin	8.86	4	0.0647	13.37	4	0.0096
All interactions	3.66	3	0.3006	4.42	3	0.2198
Trial	3948.78	2	<0.0001	4826.65	2	<0.0001
Interactions						
X_H:X_T	10.98	2	0.0041	50.92	2	<0.0001
X_H:Aspirin	1.27	2	0.5303	2.30	2	0.3159
X_T:Aspirin	3.35	2	0.1875	4.41	2	0.1102
X_H:X_T:Aspirin	0.93	1	0.3356	0.86	1	0.3530
Total Interaction	13.83	4	0.0079	54.68	4	<0.0001
Total	5293.67	9	<0.0001	7222.45	9	<0.0001

6.4.3 Absolute benefit or harm of aspirin across strata of predicted poor outcome

Benefit from treatment can vary according to patients' baseline outcome risk in the absence of treatment (Califf et al., 1997, Dorresteijn et al., 2011a, Dorresteijn et al., 2011b). With a constant relative treatment effect it is expected that the absolute reduction it the risk of six month death or dependency varies with patient prognosis. To explore this using the aspirin trial datasets patients were grouped into deciles of increasing predicted risk of poor functional outcome as estimated by the model developed for six month death or dependency in section 6.3.3. For each trial the same simple updating procedure (equation (6.1)) was applied (Table 6-11) adjusting for the observed differences in incidence of death or dependency across the three trials. The benefit of recalibration is illustrated by comparing the updated calibration metrics in Table 6-11 with the original metrics presented in Table 6-10. After updating the calibration intercepts for IST-1 and CAST were no longer negative. This is because the mean predicted risk is set equal to the incidence observed in the new data. No difference was observed for MAST-I (correction factor of 0.047) likely reflecting the similarity between the development data (UK and Italian centres) and MAST-I (an Italian study).

The IST-1 data was used to define the groups of predicted risk strata. For this analysis the predicted risk of death or dependency was categorised into deciles creating ten equal sized groups. The same grouping was then applied to the recalibrated predicted risks for CAST and MAST-I. Figure 6-8 plots the predicted risk of death or dependency if all patients were on control against the observed absolute risk difference. Each tenth was pooled using fixed effects meta-analysis.

There was no suggestion that patients with higher predicted risk of poor functional outcome experienced any greater or lesser benefit from aspirin than the overall estimate of 1% (Figure 6-8). The theoretical relationship based on the relative effect of aspirin (OR of 0.94) and the predicted risk on control appears in the plot as a dashed black line with a positive slope. Note, a global test of additivity suggested that a single OR was reasonable (P-value = 0.8232). This theoretical relationship suggests that the absolute benefit from aspirin is greater relative to the patient's risk

of poor outcome in the absence of treatment as predicted at baseline assuming a constant relative effect. Fitting a line through the deciles it is possible to test whether the theoretical relationship differs from this observed relationship. The fitted line has a slope of 0.01 (95% CI -0.04 to 0.06) were the assumption that the slope is equal to zero was not rejected (P-value = 0.6897). The theoretical relationship has a constant slope of 0.0566; an F-test was used to test whether the observed slope was significantly different from this. The P-value was 0.1655, so that the null hypothesis of *no difference* from the theoretical slope is supported by the data. Again, the exclusion of those patients randomised to streptokinase or high dose heparin made no major difference to these results (see section 6.6 Appendix A on page 177).

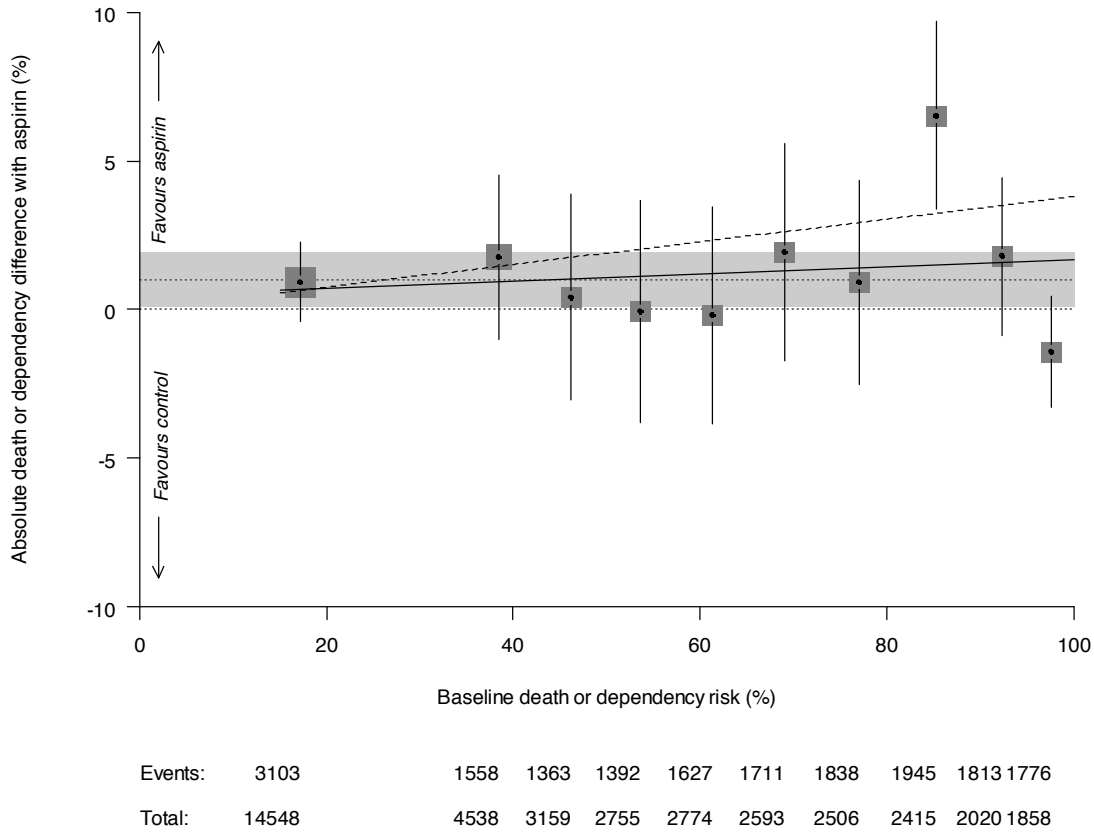


Figure 6-8 Meta-analysis of predicted risk of poor outcome (in tenths) across all three trials pooled using fixed-effects meta-analysis. Square sizes are proportional to the strata specific denominator

6.4.4 Relative effect of aspirin by predicted risk of death or dependency

Logistic regression models were used to model functional outcome as both a binary outcome as well as an ordinal outcome. The predictors used were: the logit of the predicted risk from death or dependence; aspirin and an adjustment for trial. Note that as before, the predicted risk of death or dependency (introduced in the raw linear predictor format not as a probability, centred on its mean) was allowed to interact with aspirin. The importance of trial was considerably less when compared with what was observed in the early events model (Table 6-12) were the Wald statistic for trial was about three times that of the contribution of predicted thrombotic risk. In Table 6-13 it can be seen that aspirin was significant (P-value = 0.0319 for binary LR and P-value = 0.0020 for POLR) and in neither case did the predicted risk of death or dependence interact with treatment. The effect of aspirin was estimated by both models as an OR of 0.95 (95% CI 0.92 to 1.00) and a common OR of 0.94 (95%CI 0.90 to 0.97) for the binary LR and the POLR respectively.

Table 6-13 Wald statistics for a binary LR (dead or dependent) and a proportional odds LR (across the ordinal functional outcome) modelling predicted risks from death or dependence

	Binary LR			Proportional Odds LR		
Factor	χ^2	***d.f.***	**P-value**	χ^2	***d.f.***	**P-value**
Predicted death/dependence	5119.62	2	<0.0001	7841.61	2	<0.0001
All interactions	0.11	1	0.7368	0.84	1	0.3588
Aspirin	6.89	2	0.0319	12.44	2	0.0020
All interactions	0.11	1	0.7368	0.84	1	0.3588
Trial	4380.70	2	<0.0001	5715.43	2	<0.0001
Predicted death/dependence: aspirin	0.11	1	0.7368	0.84	1	0.3588
Total	7316.80	5	<0.0001	11454.55	5	<0.0001

.

6.5 Discussion

In this chapter a stratified approach toward the treatment of acute ischaemic stroke patients with aspirin was explored. No differential treatment effects were identified amongst patients either when grouped by quarters of predicted risk from early events (thrombosis and haemorrhage) or when grouped by deciles of predicted baseline outcome risk (i.e., six month death or dependency).

Only three trial datasets were used for this analysis, however, these three trials contribute the majority of the evidence of aspirin efficacy for acute stroke (Sandercock et al., 2008). The logistical efforts made in obtaining the patient data for the purpose of this chapter were not nearly as great as that which may be anticipated by other larger studies where the obtainment of many datasets from many trials is pursued. For instance an IPDMA investigating the prognostic risk factors of foot ulceration amongst those with diabetes allowed: three to six months for *background and research*; 12 months for *data checking*; and another three to six months for *analysis and publication* (Crawford et al., 2013). Similarly, summarising the findings from the International Mission for Prognosis and Analysis of Clinical Trials in TBI (IMPACT) consortium, Mass *et al* highlighted the merging of multiple datasets into one standardised stacked dataset as one of the main struggles (Maas et al., 2013). No doubt spurred on by a degree of exasperation in this lengthy process, some have even called for the creation of a central repository within which all datasets collected and collated for IPDMA projects could be stored; thus enabling easier access to the resources created by the hard work of others (Tudur Smith et al., 2014). Therefore it is well understood that obtaining individual datasets from each of the contributing studies is frequently time consuming and potentially costly. Cost-effectiveness should be discussed in larger IPDMA studies summarising whether the added effort and time required, beyond that of a standard meta-analysis based on aggregate data, is justified. Some comments are made regarding the IPD analysis of the aspirin trial data versus the standard meta-analysis based on aggregate results.

First, it is noted that the same conclusion about the efficacy of aspirin was reached by both the aggregate meta-analysis and the IPDMA. Is this surprising? No. To

understand this consider the three most popular methods for a fixed effects meta-analysis: Peto's OR method; the Mantel-Haenszel method; and maximum likelihood estimation using logistic regression (Deeks et al., 2008). These different approaches will yield very similar results, for example, using each method on the data provided in Figure 6-4 part B the following estimates are obtained: 0.9483, 0.9484 and 0.9484 respectively. The logistic regression approach can be understood as fitting a dummy variable for treatment (associated coefficient β_1) and two dummy variables to adjust for trial (β_2 and β_3):

$$G(Y \mid Aspirin, Trial) = \beta_0 + \beta_1 Aspirin + \beta_2 Trial_{01} + \beta_3 Trial_{02} \qquad \textbf{Model 1}$$

It is now obvious that the model fit in Table 6-12 is merely an extension of this simple model (Model 1) with the inclusion of: the predicted risk of haemorrhage; the predicted risk of thrombosis; their separate interaction with treatment; their joint interaction with treatment; and the interaction of the two (δ_1, δ_2, δ_3, and δ_4).

$$\begin{aligned} G(Y \mid Aspirin, Trial, \hat{X}_H, \hat{X}_T) = {} & \beta_0 + \beta_1 Aspirin + \beta_2 Trial_{01} + \beta_3 Trial_{02} + \\ & \beta_4 \hat{X}_H + \beta_5 \hat{X}_T + \\ & \delta_1 Aspirin : \hat{X}_H + \delta_2 Aspirin : \hat{X}_T + \delta_3 \hat{X}_T : \hat{X}_H + \delta_4 Aspirin : \hat{X}_T : \hat{X}_H \end{aligned} \qquad \textbf{Model 2}$$

Interactions between aspirin and predicted risk did not improve the model fit (χ^2 of 3.66, see Table 6-12) providing insufficient evidence to reject the null hypothesis that $\delta_1 = \delta_2 = \delta_4 = 0$. The model was, however, considerably improved by the inclusion of predicted risk from early events (Model 2 Likelihood ratio χ^2 of 6359.13 versus Model 1 Likelihood ratio χ^2 of 3950.86). This had little impact on the estimate for the effect of aspirin (OR of 0.9613). It is evident then that it should in fact be *anticipated* that were there is no evidence of an interaction the IPDMA should *at least* reach the same conclusion as the aggregate result. This replication is therefore reassuring.

Second, these findings may be mistakenly interpreted as a 'failed result'. However, this would be a rather pessimistic conclusion to draw. Rothwell points out that even when no subgroup difference is identified an important finding has still been made:

this may indicate that the treatment is more generalisable than previously anticipated (Rothwell, 2007b). Such findings can therefore be used to dissuade the undertreatment of certain patient groups otherwise thought at a greater risk of harm. Based only on the aggregate data it would not have been possible to have commented on this. It is in this way that the IPDMA of aspirin trials is justifiable.

6.6 Appendix A: Sensitivity analyses

It may be the case that by including those patients on heparin and streptokinase in IST-1 and MAST-I respectively could cause a dilution of the effect of aspirin. All of the analyses undertaken in this chapter were therefore repeated excluding those randomised to receive either low dose or medium dose heparin or streptokinase.

The impact this had on the size of the trial data sizes was to reduce IST-1 from 18372 to 9197 and MAST-I from 622 to 309. Note CAST contributes the same number of patients as before. The total size of the three trials together is then 29678 patients meaning that there is a 24% reduction in the sample size contrast with the sample used throughout this chapter. What follows is a table summarising these findings (Table 6-14) and the resulting output (Table 6-15, Table 6-16, Figure 6-9 and Figure 6-10).

Excluding these patients made no qualitative change to the findings compared to those made on all data. Accuracy was likely lost due to models being fit on fewer events.

The composite early outcomes as defined in this chapter are heterogeneous. For instance the mechanism of deep venous thrombosis may differ from MI. Aspirin may have a differential effect on the reduction of the risk of occlusive arterial events and occlusive venous events. A sensitivity analysis is presented in Table 6-17 excluding venous events from the model presented in Table 6-6. This restriction made no material difference to the results.

Table 6-14 Summary of results from repeating all analyses in patients treated with aspirin or control only.

Analysis	Summary
Models development	The main impact this has is on the size of the data available for model development. This will have a impact on the estimation of parameters due to the small number of events I. Thrombotic model; effect sizes largely in the same direction, except for dysphasia which was negatively associated although not significantly. Delay from randomisation was also no longer significant. No new significant associations detected. II. Haemorrhagic model; effect sizes largely in the same direction, except for visuospatial disorder which was negatively associated although not significantly. Age was also no longer significant. No new significant associations detected. III. Death or dependence; effect sizes all in the same direction. Presence of brainstem/ cerebellar deficit now significant.
Model performance	Model performance was as poor, and with even fewer events the reliability of the estimated metrics is debatable.
16ths of predicted risk	Lower risk of haemorrhage noted. With less haemorrhagic events occurring in the MAST-I aspirin only subset the trial data contribute to more than one strata. No change in the conclusion though (see Figure 6-9)
Baseline outcome risk	Visual assessment of Figure 6-10 suggests that the best fit line through the deciles matches the theoretical slope better (P-value of no difference = 0.5423)

Table 6-15 Sensitivity analysis: Multivariable prediction models for early events and late events with imputed IST-1 development data (over 20 imputed sets)

	Thrombosis (153/2266)		Haemorrhage(18/2239)		Dead/dependent (1608/2266)	
Variable	**Coef (SE)**	**P-value**	**Coef (SE)**	**P-value**	**Coef (SE)**	**P-value**
Intercept	-5.119 (0.944)	<0.0001	-4.650 (2.540)	0.0671	2.920 (1.870)	0.1184
Age (per 10 years)						
Age	0.261 (0.092)	0.0046	-0.030 (0.239)	0.9004	-0.758 (0.380)	0.0459
Age^2	-	-	-	-	0.092 (0.029)	0.0014
SBP (per 10 mmHg)						
SBP	-0.023 (0.032)	0.4779	-0.055 (0.093)	0.5522	-0.341 (0.192)	0.0762
SBP^2	-	-	-	-	0.011 (0.006)	0.0699
Delay from randomisation (hours)	0.001 (0.007)	0.8770	-0.016 (0.020)	0.4092	0.003 (0.004)	0.4445
Sex (Male)	0.179 (0.176)	0.3090	0.723 (0.525)	0.1683	-0.346 (0.110)	0.0017
Visible infarct on CT	0.278 (0.188)	0.1382	0.789 (0.515)	0.1252	0.097 (0.124)	0.4315
Conscious (Drowsy/Coma vs. Alert)	-0.105 (0.230)	0.6468	0.134 (0.625)	0.8304	1.559 (0.228)	<0.0001
Presence of atrial fibrillation	0.304 (0.206)	0.1408	0.560 (0.602)	0.3520	0.091 (0.162)	0.5764
At least one or more of arm/leg deficits	0.570 (0.337)	0.0906	0.792 (1.070)	0.4591	0.854 (0.159)	<0.0001
Presence of dysphasia	-0.014 (0.182)	0.9395	-0.983 (0.558)	0.0784	0.486 (0.113)	<0.0001
Presence of hemianopia	0.307 (0.264)	0.2447	0.401 (0.665)	0.5463	0.903 (0.205)	<0.0001
Presence of visuospatial disorder	0.054 (0.255)	0.8316	-0.311 (0.673)	0.6443	0.631 (0.190)	0.0009
Presence of brainstem/ cerebellar deficit	0.300 (0.294)	0.3075	-0.533 (1.065)	0.6164	-0.344 (0.173)	0.0460

Table 6-16 Sensitivity analysis: model performance in IST-1 evaluation split. Measures pooled over 20 multiply imputed data

Performance measure	Thrombosis (199/2332)	Haemorrhage (19/2360)	Dead/dependent (1296/2332)
AUROCC	0.59 (0.55 to 0.63)	0.51 (0.39 to 0.64)	0.77 (0.75 to 0.79)
Calibration			
Intercept	0.34 (0.26 to 0.41)	-0.09 (-0.32 to 0.14)	-0.75 (-0.78 to -0.71)
Slope	0.62 (0.45 to 0.78)	0.04 (-0.24 to 0.32)	0.89 (0.86 to 0.93)

Table 6-17 Sensitivity analysis: Multivariable prediction model for 14 day thrombotic events excluding DVTs (260/4504)

Variable	Coefficient (SE)	OR (95% CI)	P-value
Intercept	-4.529 (0.697)	-	<0.0001
Age (per decade)	0.231 (0.069)	1.26 (1.10 to 1.44)	0.0008
SBP (per 10 mmHg)	-0.030 (0.025)	0.97 (0.92 to 1.02)	0.2191
Delay from randomisation (hours)	-0.009 (0.005)	0.99 (0.98 to 1.00)	0.0976
Sex (Male)	0.244 (0.134)	1.28 (0.98 to 1.66)	0.0673
Visible infarct on CT	0.373 (0.143)	1.45 (1.10 to 1.92)	0.0089
Conscious (Drowsy/Coma vs. Alert)	-0.031 (0.170)	0.97 (0.70 to 1.35)	0.8539
Atrial fibrillation	0.127 (0.161)	1.14 (0.83 to 1.56)	0.4293
At least one or more of arm/leg deficits	0.303 (0.227)	1.35 (0.87 to 2.11)	0.1831
Presence of dysphasia	0.104 (0.139)	1.11 (0.85 to 1.46)	0.4545
Presence of hemianopia	0.043 (0.200)	1.04 (0.71 to 1.54)	0.8318
Presence of visuospatial disorder	0.106 (0.188)	1.11 (0.77 to 1.61)	0.5726
Presence of brainstem/ cerebellar deficit	0.292 (0.213)	1.34 (0.88 to 2.03)	0.1706

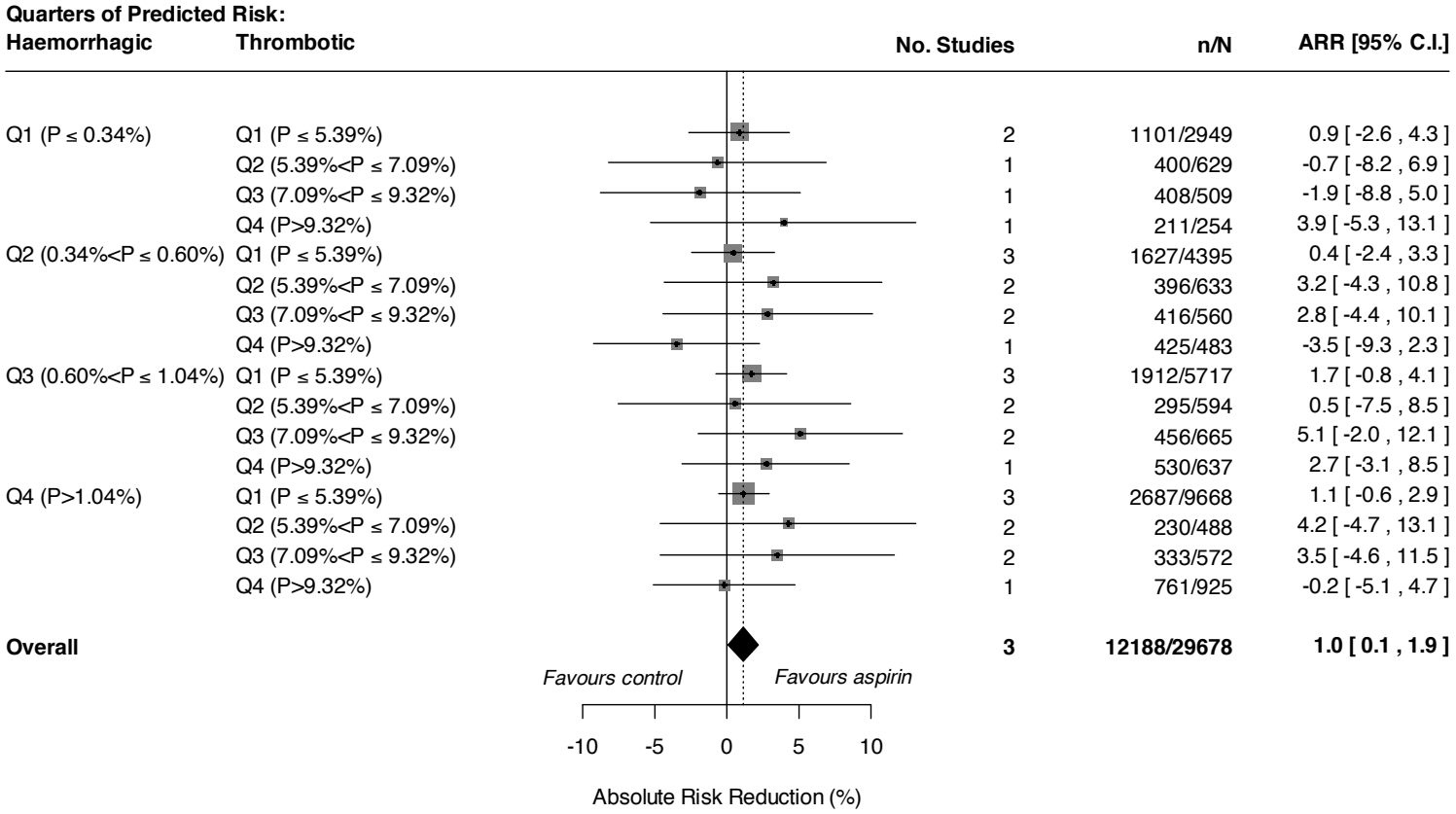


Figure 6-9 Sensitivity analysis: pooled estimates of ARR in poor outcome. Q1 to Q4 denote quarters of risk. Square sizes are proportional to the strata specific denominator

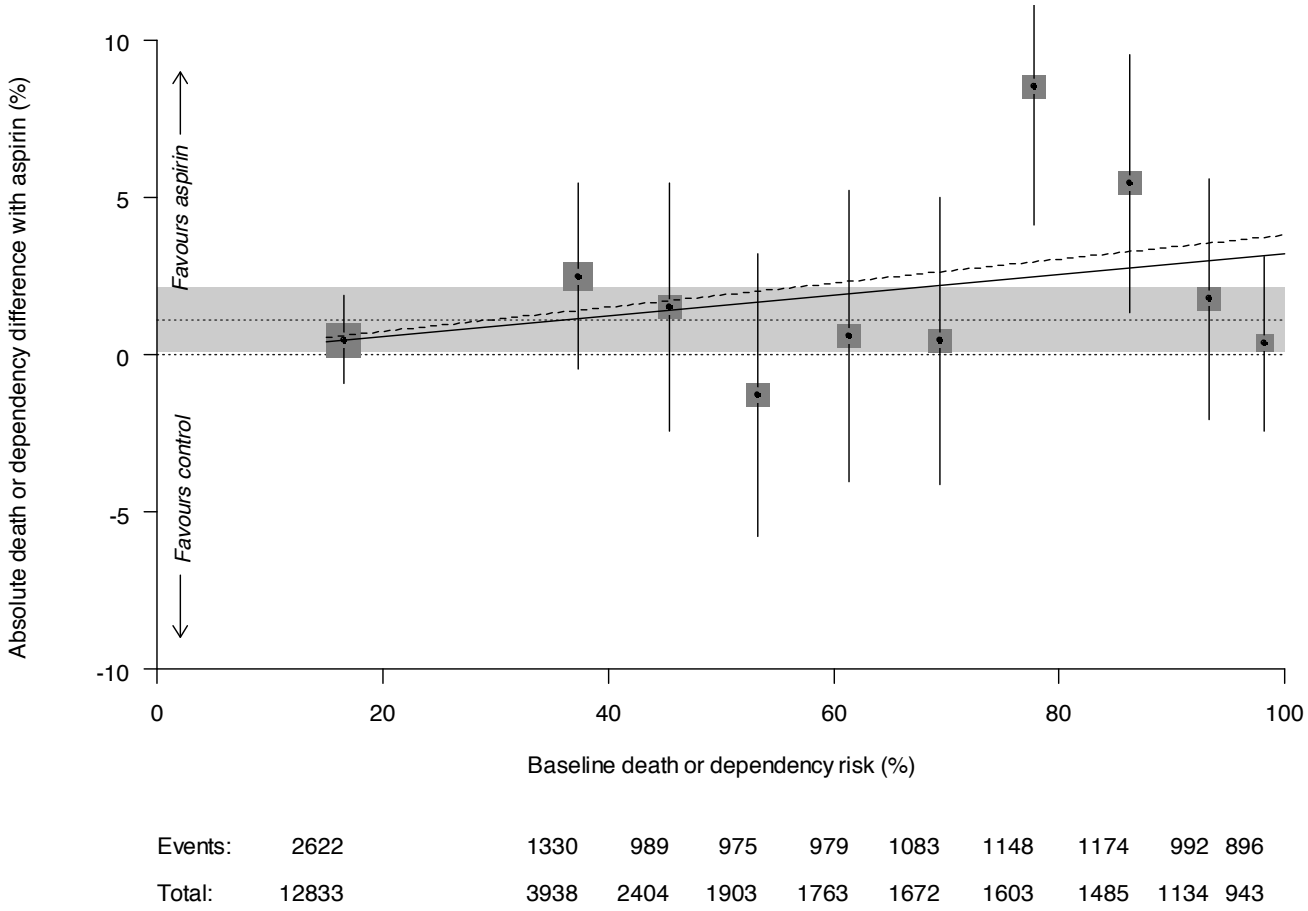


Figure 6-10 Sensitivity analysis: meta-analysis of predicted risk of poor outcome (in tenths) across all three trials pooled using fixed-effects meta-analysis. Square sizes are proportional to the strata specific denominator

Chapter 7: Benefits and harms from iv-rtPA in acute ischaemic stroke

Background and summary

It is unclear whether those at highest risk of symptomatic intracranial haemorrhage or poor functional outcome after acute ischaemic stroke benefit from treatment with intravenous rtPA. Using the IST-3 data, there was no evidence to suggest that those at a high predicted risk of intracranial haemorrhage or poor functional outcome should avoid rtPA.

7.1 Introduction

Thrombolytic treatments are a class of drugs used to dissolve blood clots and restore normal blood flow. Tissue plasminogen activator (tPA) is a naturally occurring thrombolytic which catalyses the conversion of inactive plasminogen into active plasmin, an enzyme responsible for clot breakdown (Collen and Lijnen, 2009). Recombinant tPA is a manufactured form of tPA (also known as rtPA or alteplase) and is one of the few known effective treatments for acute ischaemic stroke (Adams et al., 2007, Wardlaw et al., 2013).

Between 1995 and 2012 the efficacy of rtPA was assessed across 12 separate trials randomising a total of 7012 acute ischaemic stroke patients to receive rtPA or a comparator (Wardlaw et al., 2012a). Two of these studies concluded that rtPA significantly reduced the odds of a poor outcome over that of control (National Institute of Neurological Disorders and Stroke, NINDS trial and the second European-Australasian Acute Stroke Study, ECASS-III) (The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group, 1995, Hacke et al., 2008). A meta-analysis conducted by Wardlaw *et al*. showed that rtPA was associated with a significant increase in the odds of a favourable outcome with mRS≤2 (OR of 1.17 with 95%CI of 1.06 to 1.29) (Wardlaw et al., 2012a).

Exactly who benefits has been an area of ongoing research (The Stroke Thrombolysis Trialists' Collaborative Group, 2013, Sandercock et al., 2011). One of the harmful side-effects of treatment with rtPA is an increase in the odds of Symptomatic Intracranial Haemorrhage (SICH) with an OR of 3.72 (95%CI 2.98 to 4.64). A bleed into the brain is a catastrophic event which ultimately results in severe disability or death. Physicians are justifiably cautious when prescribing rtPA; this is reflected in the licence for its use. The current European licence for rtPA restricts its use to patients with: SBP <185mmHg; blood glucose ≤400mg/dl (22mmol/l); patients with small cerebral infarction; and moderate stroke severity (an NIHSS >4, but NIHSS <25) (The European Stroke Organisation Executive Committee, 2008). A key issue is the impact that delay from stroke onset to the administration of treatment has on the expected benefit. Evidence has suggested that an interaction exists between rtPA and the onset time to treatment (OTT) which favours the earliest administration of treatment to those eligible patients so as to ensure the largest beneficial effect from treatment (Lees et al., 2010). The current window of OTT for the North American licence is restricted to three hours from onset, whilst in Europe a maximum of four and a half hours is permitted.

It is possible that those at high risk from SICH post rtPA or low risk from poor functional outcome may be harmed by treatment. A formalised strategy based on predictions is therefore a reasonable strategy to explore (Kent et al., 2010). Indeed prediction made by clinical prediction models may do better than the informal judgements made by clinicians. In this chapter, the use of clinical prediction models for SICH post rtPA will be reviewed using data from the third International Stroke Trial (IST-3): the largest trial of rtPA in acute ischaemic stroke patients. A framework for treating patients with rtPA will be explored based on their predicted risk of post rtPA SICH or poor functional outcome. The aims of this chapter are threefold: (i) to evaluate the performance of existing models; (ii) to develop a new model in IST-3 adopting a more rigorous statistical methodology; and (iii) to investigation whether those at high predicted risk of SICH or low risk of poor functional outcome suffer any overall harm from treatment with rtPA.

7.2 Data and definitions

7.2.1 The third International Stroke Trial (IST-3)

The third International Stroke Trial (IST-3) was a pragmatic, international, multicentre, randomised-controlled, open-treatment trial of intravenous rtPA in acute ischaemic stroke (The IST-3 collaborative group, 2012a). The active treatment was given at a dose of 0.9 mg/kg with 10% administered as a single dose with the remaining amount given over a one hour period. All patients had planned follow-up of six months – though some contributing countries had a longer follow-up period of 18 months as part of a pre-planned secondary objective assessing the long term impact of rtPA (The IST-3 collaborative group, 2013).

The IST-3 trial was designed to assess the benefits and harms of rtPA within a wider patient population than had been previously tested for thrombolysis treatment; this meant that the majority of those recruited to IST-3 did not meet the license criteria for treatment with rtPA (Sandercock et al., 2011). No upper age limit was set and a wider window of onset time to treatment was allowed of up to six hours. The primary outcome was the proportion of those alive and independent at six months as measured on the Oxford Handicap Scale (OHS). This was defined both as a dichotomy with an OHS≥3 as a poor outcome as well as on the original ordinal scale (The IST-3 collaborative group, 2012b). Minimisation was used to randomise those recruited. This is a dynamic method of randomisation which ensures balance between two treatment arms with respect to pre-specified patient characteristics (Senn, 2007). The minimisation algorithm had an allocation probability of 0.8 and balanced on: region; age; sex; NIHSS; treatment with antiplatelet in previous 48 hours; and stroke clinical syndromes. Patients were recruited from May 2000 till July 2011 from 12 countries and 156 centres. A total of 3035 patients were randomised to receive recombinant tissue plasminogen activator (rtPA) plus standard care or standard care alone. Patients were enrolled into the trial using the uncertainty principle meaning that both the treating physician and the patient (or a proxy) felt that the benefit of rtPA was *promising but unproven*. Patients were eligible provided that: they had symptoms and signs of clinically definite acute stroke; that the onset time of the stroke was known; randomisation to trial treatment could be started

within 6 hours of the index stroke; and that an intracranial haemorrhage or mimic stroke (e.g., cerebral tumor) could be reliably excluded using a scan (e.g., on CT or MRI). Complete six-month follow-up was achieved. There was no significant difference in the effect of treatment on the binary categorisation of the OHS (with a P-value of 0.181) though there was a significant shift which favored treatment with rtPA over standard therapy alone (with a P-value of 0.001).

The reporting of the IST-3 was completely transparent with the study protocol (Sandercock et al., 2011) and the statistical analysis plan (The IST-3 collaborative group, 2012b) published prior to any analysis.

7.2.2 Defining patient characteristics and outcomes

Baseline data were collected either through a voice-activated telephone line, or else using a secure website system. Biologically implausible values were flagged up as and when they arose. Specific attention was given to the way in which brain scans were carried out. This is discussed later in this chapter. Symptomatic Intracranial Haemorrhage (SICH) by seven days was defined in the IST-3 protocol as:

> "… clear evidence of significant intracranial haemorrhage on the post-randomisation scan (or autopsy if not re-scanned and death occurs after 7 days). Significant haemorrhage is present on any post-randomisation scan if the expert reader gives any response to Question 22 other than a blank value or 'Petechial haemorrhage' (i.e. significant HTI, parenchymal haematoma, etc) AND a response to Question 23 of 'yes', indicating that haemorrhage is a major component of the IST-3 lesion (or is remote from the lesion and likely to have contributed significantly to the burden of brain damage). This event includes clinical events described as a recurrent stroke within 7 days, in which the recurrent stroke was confirmed to be due to an intracranial haemorrhage."

Functional outcome was recorded at six months for each patient using the Oxford Handicap Scale (OHS): a measure grading patients on an ordinal disability scale. The OHS ranges from 0 (recovered) in unit intervals to 6 (dead). The binary classification of functional outcome was defined with a dichotomisation at OHS≥3.

7.2.3 Sensitivity Analyses

It was noted in the introduction that the size of the effect of rtPA is strongly related to the delay from stroke onset to receipt of treatment (Lees et al., 2010). The patients recruited as part of IST-3 could be randomised as late as six hours from stroke onset. All the analyses conducted in this chapter were therefore repeated amongst those receiving rtPA within four and a half hours as well as those treated after four and a half hours. Additionally, there are various definitions for haemorrhage. One alternative definition of haemorrhage was explored: parenchymal haemorrhage. Finally, a per-protocol analysis was conducted, restricting analysis to just those that adhered to the protocol specified by IST-3.

7.3 Predicting post rtPA events

A number of pre-existing models are now introduced. An extensive evaluation of their performance in the IST-3 dataset follows in section 7.3.4.

7.3.1 Selection of pre-existing models

A literature search was conducted from inception to 25th January 2013 (Dr Whiteley) for studies which had developed multivariable clinical prediction models for the prediction of intracranial haemorrhage or poor functional outcome in acute ischemic stroke patients treated with rtPA. Models were eligible if developed within a cohort of acute ischemic stroke patients, all of whom had been treated with rtPA, and reported a prediction model either as a score, or a model with constant and weighting for each of the covariates. Studies were identified using: an electronic search strategy; reference lists of relevant studies; forward searches from relevant studies with Google Scholar; and from personal files. The literature search identified 797 publications of which 11 studies were relevant, reporting the development of 12 models, nine of which could be applied using the IST-3 dataset (see Table 7-1 for an account of the variables included in each).

All were derived as multivariable binary logistic regression models except for the HAT and the SPAN-100 scores which were derived using existing knowledge with no modelling. Most (seven out of nine) were presented as a point score.

7.3.2 Models for SICH

7.3.2.1 The HAT score

The Haemorrhage after Thrombolysis (HAT) score was developed by systematically reviewing risk factor studies (Lou et al., 2008). The authors used a selection criterion requiring that at least two studies supported the importance of a clinical risk factor associated with SICH post rtPA. Points where assigned based on the strength of the odds ratios reported. The authors evaluated their model in the NINDS trials group which suggested good discriminative ability with AUROCC values ranging from 0.68 to 0.75 depending on how haemorrhage was defined. A recent study supported this finding with 0.70, 0.69 and 0.73 with SICH as defined by the National Institute of Neurological Disorders and Stroke (NINDS), the second European-Australasian Acute Stroke Study (ECASS II) and the Safe Implementation of Thrombolysis in Stroke-Monitoring Study (SITS-MOST) respectively (Sung et al., 2013). Strbian *et al* found a moderate level of discrimination in their data using the same definitions of SICH: 0.65, 0.65 and 0.64, whilst an evaluation by Cucchiara *et al* suggested a more conservative AUROCC of 0.59 (Cucchiara et al., 2011, Strbian et al., 2014).

7.3.2.2 The SEDAN score

The Symptomatic Intracranial Hemorrhage after Stroke Thrombolysis (SEDAN) score was developed on 974 ischaemic stroke patients treated with rtPA within four and a half hours of stroke onset (Strbian et al., 2012a). Patients were recruited consecutively at several centers in Helsinki, Finland. Univariate screening (P-value <0.1) followed by backward stepwise selection (P-value <0.1) was used for predictor selection. The final score included five clinical risk factors. The authors presented an external evaluation which suggested good discrimination (AUROCC of 0.77) though a separate evaluation indicated a more moderate level of discrimination with an AUROCC of 0.60 where SICH was defined by SITS-MOST and 0.66 for ECASS-II (Mazya et al., 2013). Another study supported a higher level of discrimination that was little affected by the definition of SICH with AUROCC values of 0.69, 0.70 and 0.69 with SICH as defined in NINDS, ECASS II and SITS-MOST respectively (Strbian et al., 2014).

7.3.2.3 The SITS score

The Safe Implementation of Treatments in Stroke (SITS) SICH risk score was developed within a large multinational registry of centers treating ischaemic stroke patients with thrombolysis (Mazya et al., 2012). A random 50:50 split meant that half were used in model development (15814) and the rest to evaluate model performance (15813). The score included nine statistically significant risk factors after a process of univariate screening (P-value ≤0.10) and screening for significance in a multivariable model fit (P-value <0.05). Discrimination was good with an AUROCC of 0.70. Sung *et al*. suggested that the SITS score achieved moderate discrimination with AUROCC values of 0.62, 0.61 ad 0.68 with SICH defined respectively according to NINDS, ECASS II and SITS-MOST (Sung et al., 2013). Strbian *et al*. found similar levels with AUROCC values of 0.61, 0.64 and 0.67 using the same definitions for SICH (Strbian et al., 2014).

7.3.2.4 The GRASPS score

The Glucose Race Age Sex Pressure Stroke Severity (GRASPS) score used data from the: Get with the Guidelines–Stroke (GWTG-Stroke) register (Menon et al., 2012). The authors used 10242 patients and implemented a 70:30 split, developing the model on the majority of the data and using the remaining proportion for internal evaluation. Sung *et al* tested the discriminative ability of the GRASPS score in their prospectively collected hospital data and found AUROCC values of 0.62, 0.61 and 0.63 when SICH was defined as according to NINDS, ECASS-II and SITS-MOST respectively whilst Strbian *et al* found an AUROCC of 0.67 irrespective of the way SICH was defined (Strbian et al., 2014, Sung et al., 2013).

7.3.2.5 The SPAN-100 score

The Stroke Prognostication using Age and NIH Stroke Scale (SPAN) index combined two common determinants of stroke prognosis as a rationale for predicting ICH and favorable outcome (mRS≤1) after treatment with thrombolysis (Saposnik et al., 2013b). The authors found that amongst those given rtPA, ICH rates for *SPAN + ve* patients were higher than those *SPAN –ve* patients (42% vs. 12%). Subsequent external evaluations of the SPAN-100 index have been disappointing with Sung *et al* reporting AUROCC values of 0.56, 0.55 and 0.57 with SICH defined as according to

NINDS, ECASS-II and SITS-MOST respectively with similar findings made by Strbian *et al* with AUROCCs of 0.55, 0.56 and 0.56 (Sung et al., 2013, Strbian et al., 2014).

7.3.3 Models for poor functional outcome

7.3.3.1 The Stroke-TPI score

The Stroke Thrombolytic Predictive Instrument (TPI) used randomised clinical trial data from five trials investigating the use of rtPA within six hours of stroke onset (Kent et al., 2006). The authors developed models for the prediction of mRS≤1 and mRS≥5. Predictor selection was carefully considered with the authors acknowledging the risk of overfitting and the inclusion of variables difficult to record in practice. A core set of important risk factors where used. Discriminative performance in an independent dataset was good (AUROCC of 0.77) as was calibration. An external evaluation by McMeekin *et al*. demonstrated the same level of discrimination (McMeekin et al., 2012).

7.3.3.2 The DRAGON score

The (hyper)Dense middle cerebral artery sign or early infarct signs on admission CT head scan, prestroke modified Rankin Scale score >1, Age, Glucose level on admission, OTT, and NIHSS (DRAGON) score was developed and evaluated on 1319 consecutive rtPA treated acute ischaemic stroke patients seen at hospital in Helsinki, Finland (Strbian et al., 2012b). Variables were selected *a priori* and retained regardless of statistical significance. However, an additional 15 risk factors were not included since no statistical association was identified. Performance in external evaluation (333 patients) was good with an AUROCC of 0.80 (95% CI 0.74 to 0.86). A subsequent study supported this discriminative performance (Strbian et al., 2013).

7.3.3.3 The THRIVE score

The Totaled Health Risks in Vascular Events (THRIVE) score was developed using data from two multicenter trial, the Mechanical Embolus Removal in Cerebral Ischemia (MERCI) and Multi MERCI trials (Flint et al., 2010). A multivariable

model was developed, although it was not clear how the authors chose risk factors for inclusion.

7.3.3.4 A NIHSS/age model

König *et al*. developed a model within the Virtual International Stroke Trial Archive (VISTA) data base (König et al., 2008, Weimar et al., 2002). This model was used to explore subgroup effects in the IST-3 trial. It was developed to predict Barthel Index (BI, a measure of functional independence) score of less than 95 or death by three months.

Table 7-1 Models for the prediction of post rtPA SICH and poor functional outcome.

	Post rtPA SICH					Post rtPA poor outcome			
Variables	HAT	SEDAN	SITS	GRASPS	SPAN-100	Stroke-TPI	DRAGON	THRIVE	König *et al*
NIHSS	•	•	•	•	•	•	•	•	•
Glucose	•	•	•	•		•	•	•	
Age		•	•	•	•	•	•	•	•
Visible infarct on CT	•	•				•	•		
Hyperdense artery on CT		•					•		
Systolic blood pressure			•	•					
Prior hypertension			•					•	
Use of antiplatelets			•						
Weight			•						
Delay to rtPA			•				•		
Sex				•					
Ethnicity				•					
Prior disability							•		

7.3.4 Model performance in IST-3

All assessments of model performance were undertaken within those treated with rtPA in the IST-3 dataset. No imputation of missing data was undertaken. Similar risk factors were included in each of the models irrespective of the outcome they were developed to predict. It was anticipated that discriminatory ability would be similar whether using a model original derived to predict SICH to predict poor outcome and vice versa. Note that the AUROCC measure is a rank based measure and therefore unaffected by differences in baseline risk (i.e., contrasting the incidence of SICH post rtPA to poor functional outcome).

7.3.4.1 Cohort comparability

Patient baseline information was extracted from each of the development cohorts where possible from the original published articles (see Table 7-2, Table 7-3 and Table 7-4). Additional baseline information recorded but not included in the final model was also extracted in a bid to understand differences between the development cohorts and the IST-3 cohort. No baseline information could be extracted for the THRIVE model or the König model (Flint et al., 2010, König et al., 2008).

On average those recruited to IST-3 were older than most of those included in the development cohorts with a typically longer onset time to treatment, except for Stroke-TPI patients who had similar onset times to treatment. Stroke severity as measured using the NIHS score was broadly comparable, bar one exception, indicating that those in the SPAN-100 evaluation cohort suffered more severe strokes on average. Also, where reported, IST-3 tended to have a larger proportion of positive findings amongst brain imaging variables.

Table 7-2 Baseline characteristics of derivation cohorts for post rtPA SICH models (SEDAN and GRASPS) and one evaluation cohort (HAT) contrast with the IST-3 cohort. Rows in italics highlight variables not included in the final model.

Risk factors	Summary	IST-3	HAT	Summary	IST-3	SEDAN	Summary	IST-3	GRASPS
NIHS Score	<15 (%)	64	52	Median (IQR)	11 (12)	10 (9)	<5 (%)	20	18
	15-20 (%)	19	31	-		-	6-10 (%)	28	28
	≥20 (%)	17	17	-		-	11-15 (%)	20	21
	-	-	-	-	-	-	16-20 (%)	18	18
	-	-	-	-	-	-	>20 (%)	14	15
Glucose level (mg/dl)	>200 or DM (%)	6	3	Median (IQR)	7 (2)	7 (2)	<100 (%)	17	16
	-	-	-	-	-	-	100-149 (%)	56	60
	-	-	-	-	-	-	≥150 (%)	18	23
	-	-	-	-	-	-	Missing	9	-
Age	*Mean (SD)*	*78 (12)*	*71 (17)*	Median (IQR)	81 (14)	70 (17)	≤60 (%)	10	27
	-	-	-	-	-	-	61-70 (%)	12	20
	-	-	-	-	-	-	71-80 (%)	23	25
	-	-	-	-	-	-	>80 (%)	54	28
CT appearance of infarct	None (%)	61	77	Present (%)	41	34	-	-	-
	<1/3MCA (%)	24	20	-	-	-	-	-	-
	≥1/3MCA (%)	15	3	-	-	-	-	-	-
CT Hyper dense artery	-	-	-	Present (%)	25	19	-	-	-

Table 7-2 Continued from previous page

Risk factors	Summary	IST-3	HAT	Summary	IST-3	SEDAN	Summary	IST-3	GRASPS
Systolic BP (mmHg)	-	-	-	*Median (IQR)*	*156 (31)*	*156 (31)*	<120 (%)	6	8
	-	-	-	-	-	-	120-149 (%)	33	33
	-	-	-	-	-	-	150-179 (%)	44	39
	-	-	-	-	-	-	≥180 (%)	17	20
Prior hypertension	-	-	-	*Present (%)*	*64*	*59*	-	-	-
Delay to rtPA (mins)	-	-	-	*Median (IQR)-*	*231 (117)*	*120 (68)*	-	-	-
Sex	-	-	-	-	-	-	Male (%)	48	51
Ethnicity	-	-	-	-	-	-	Asian (%)	NA	2

ABBREVIATIONS: DM - Diabetes Mellitus; IQR - Interquartile Range; SD - Standard Deviation; NIHS - National Institutes of Health Stroke; CT - Computerised Tomography; MCA - Middle Cerebral Artery; BP - Blood Pressure; mmHg - millimeter of mercury; mg/dl – milligram per deciliter

Table 7-3 Baseline characteristics of an evaluation cohort for the SPAN-100 model and the derivation cohort for the SITS model contrast with the IST-3 cohort. Rows in italics highlight variables not included in the final model. Note that for these two models reported characteristics are split by defined strata, thus IST-3 has been summarised accordingly.

	Summary		IST-3		SPAN-100	Summary		IST-3		SITS
Risk factors		**SPAN negative**	**SPAN positive**	**SPAN negative**	**SPAN positive**		**SICH**	**No SICH**	**SICH**	**No SICH**
NIHS Score	Mean (SD)	9 (5)	20 (5)	14 (7)	24 (4)	Median (IQR)	15 (11)	11 (11)	14 (10)	12 (10)
Glucose level (mg/dl)	*Mean (SD)*	*130 (46)*	*133 (39)*	*150 (79)*	*137 (43)*	Median (IQR)	126 (54)	126 (36)	127 (56)	118 (39)
Age	Mean (SD)	74 (12)	87 (5)	65 (11)	81 (5)	Median (IQR)	81 (12)	81 (14)	74 (12)	69 (16)
Visible infarct on CT	*Present (%)*	*38*	*50*	*29*	*52*	*Present (%)*	*54*	*40*	*37*	*22*
Systolic BP (mmHg)	*Mean (SD)*	*156 (24)*	*155 (24)*	*158 (28)*	*157 (25)*	Median (IQR)	160 (31)	155 (31)	160 (25)	150 (30)
Prior hypertension	*Present (%)*	*64*	*66*	*65*	*80*	Present (%)	62	65	78	63
Use of antiplatelets	*Aspirin (%)*	*44*	*53*	*34*	*40*	Aspirin (%)	66	46	50	31
	Heparin (%)	*1*	*2*	*2*	*2*	Clopidogrel (%)	5	5	5	4
Weight (kg)	*Mean (SD)*	*74 (15)*	*66 (12)*	*78 (19)*	*71 (15)*	Median (IQR)	70 (15)	70 (18)	79 (20)	75 (18)
Delay to rtPA (mins)	*Mean (SD)*	*242 (71)*	*201 (70)*	*50 (65)*	*49 (33)*	Median (IQR)	222 (110)	231 (117)	150 (56)	145 (53)

ABBREVIATIONS: IQR - Interquartile Range; SD - Standard Deviation; NIHS - National Institutes of Health Stroke; CT - Computerised Tomography; BP - Blood Pressure; mmHg - millimeter of mercury; mg/dl – milligram per deciliter

Table 7-4 Baseline characteristics of the derivation cohort for the Stroke-TPI model and the DRAGON score contrast with the IST-3 cohort. Rows in italics highlight variables not included in the final model.

Risk factors	Summary	IST-3	Stroke-TPI	Summary	IST-3	DRAGON
NIHS Score	Median (IQR)	11 (12)	12 (9)	Median (IQR)	11 (12)	9 (9)
Glucose level	Median (IQR) mmol/l	7 (2)	7 (3)	Median (IQR)	7 (2)	7 (2)
Age	Mean (SD)	78 (12)	66 (11)	Median (IQR)	81 (14)	69 (17)
CT appearance of infarct	-	-	-	Present (%)	41	31
CT Hyper dense artery	-	-	-	Present (%)	25	*18*
Total ASPECTS	Median (IQR)	12 (2)	NA	-	-	-
Systolic BP (mmHg)	*Mean (SD)*	*156 (24)*	*153 (20)*	Median (IQR)	*156 (31)*	156 (31)
Prior hypertension	*Present (%)*	*64*	*59*	Present (%)	64	60
mRS score >1, prestroke	-	-	-	Present (%)	23	6
Delay to rtPA (mins)	*Mean (SD)*	*231 (73)*	*235 (135)*	Median (IQR)	*231 (117)*	118 (70)
Sex	*Male (%)*	*48*	*55*	*Male (%)*	*48*	*55*

ABBREVIATIONS: IQR - Interquartile Range; SD - Standard Deviation; NIHS - National Institutes of Health Stroke; CT - Computerised Tomography; BP - Blood Pressure; mmHg - millimeter of mercury; mg/dl – milligram per deciliter; kg – Kilogram; ASPECTS - The Alberta Stroke Program Early CT score

7.3.4.2 Calibration

Ideally the predicted risk of post rtPA SICH or poor functional outcome would equal that which is actually observed. Calibration graphs and estimates for the slope and intercept of the calibration line were used to explore how well each score predicted patient risk in IST-3. No re-calibration was undertaken for the analysis presented in Table 7-5, Figure 7-1 and Figure 7-2, meaning that the point of reference for each score was the risk originally observed in development or else a previous evaluation (as mentioned above in the case of HAT and SPAN-100).

As pointed out in Chapter 4, models presented as risk scores offer a restricted insight into the calibration of a model. With a small number of identifiable predicted risk groups the estimated calibration metrics will be subject to poor accuracy when contrast with the original model (i.e., beta coefficients and intercept available). In this respect the figures shown below for those score models are of more value than the estimates in the tables.

Models for SICH post rtPA were poorly calibrated in the IST-3 data. All models had a slope less than one suggesting that the weights associated with the risk factors per score were optimistically large. The HAT, SEDAN and SPAN-100 scores each had negative intercepts indicating a systematic over prediction of SICH risk post rtPA due to a higher observed rate of SICH in the development cohorts. The SITS and GRASPS scores had positive intercepts suggesting a systematic under prediction of the SICH risk.

Models for poor functional outcome (OHS≥3) were well calibrated amongst the IST-3 patients. The Stroke-TPI score was developed to predict a more severe functional outcome (mRS≥5, a common variant of OHS) so a systematic under prediction of mRS≥2 could be anticipated. If the outcome was fixed to OHS≥5 the intercept dropped to 1.22 (95% CI 1.09 to 1.34) and for an OHS of 6 the intercept dropped to 0.37 (95% CI 0.23 to 0.51). The DRAGON, THRIVE and König scores all had low calibration intercepts suggesting small systematic under prediction. Calibration slopes were close to one indicating that the weights associated with each risk factor where adequate for the IST-3 trial cohort.

Table 7-5 Calibration statistics of the five prediction scores for risk of SICH post rtPA and the risk of poor functional outcome (OHS≥3) post rtPA in the IST-3 dataset (N = 1515).

Risk Score	n/N	Intercept (95% CI)	Slope (95% CI)
SICH models			
HAT	87/1365	-0.29 (-0.52 to -0.05)	0.39 (0.21 to 0.57)
SEDAN	87/1365	-0.46 (-0.69 to -0.24)	0.53 (0.27 to 0.78)
SITS	85/1357	0.98 (0.76 to 1.20)	0.76 (0.41 to 1.11)
GRASPS	87/1365	0.28 (0.06 to 0.50)	0.62 (0.30 to 0.94)
SPAN-100	102/1507	-1.35 (-1.55 to -1.14)	0.36 (0.11 to 0.61)
Poor functional outcome models			
Stroke-TPI	856/1365	2.49 (2.37 to 2.62)	0.99 (0.87 to 1.11)
DRAGON	855/1363	0.20 (0.07 to 0.33)	0.96 (0.84 to 1.08)
THRIVE	955/1504	0.26 (0.15 to 0.37)	1.07 (0.91 to 1.23)
König model	957/1507	0.03 (-0.10 to 0.15)	0.84 (0.74 to 0.94)

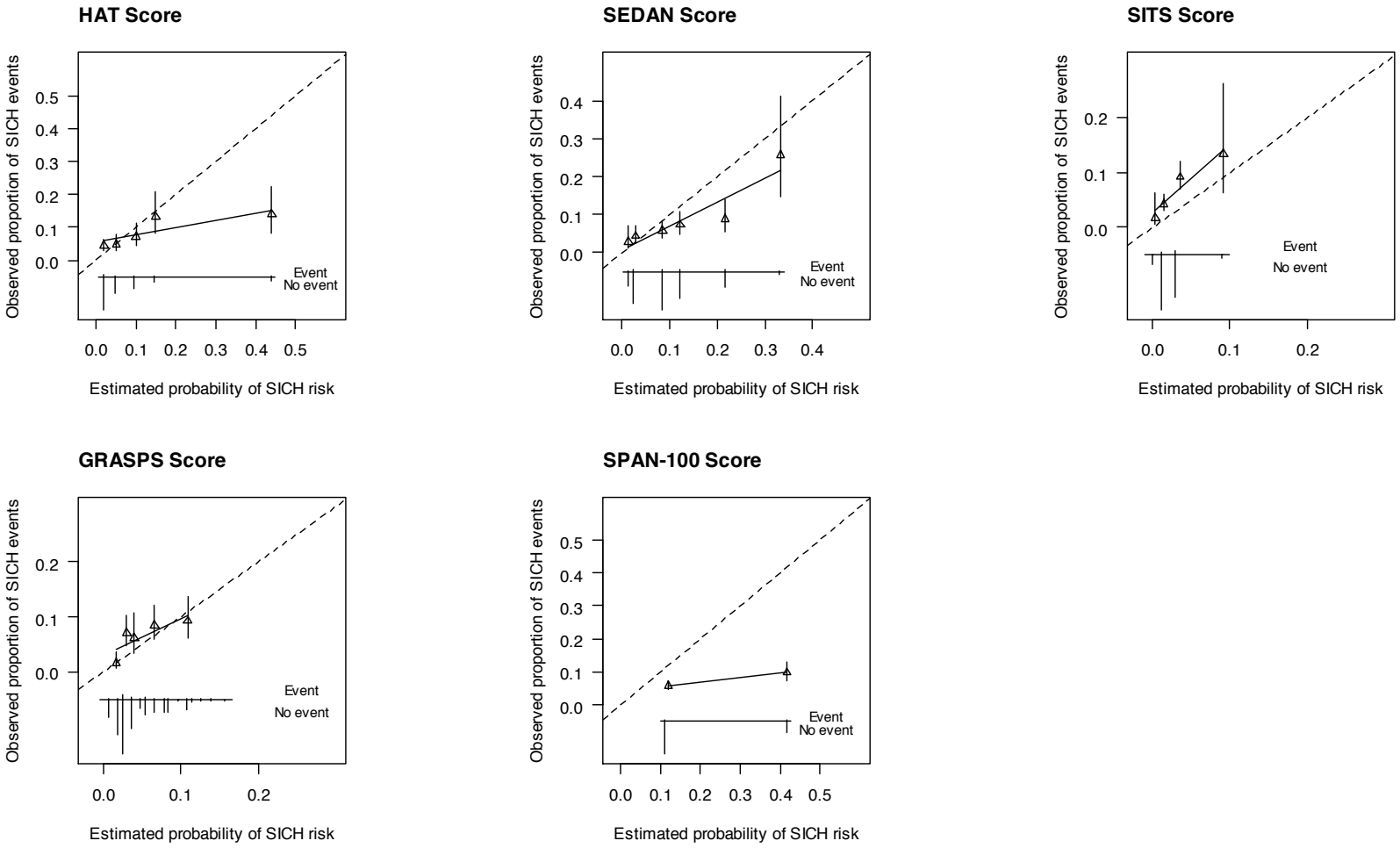


Figure 7-1 Calibration graphs for risk of SICH. Each plot shows the predicted risk groups identified by the model versus the observed frequencies.

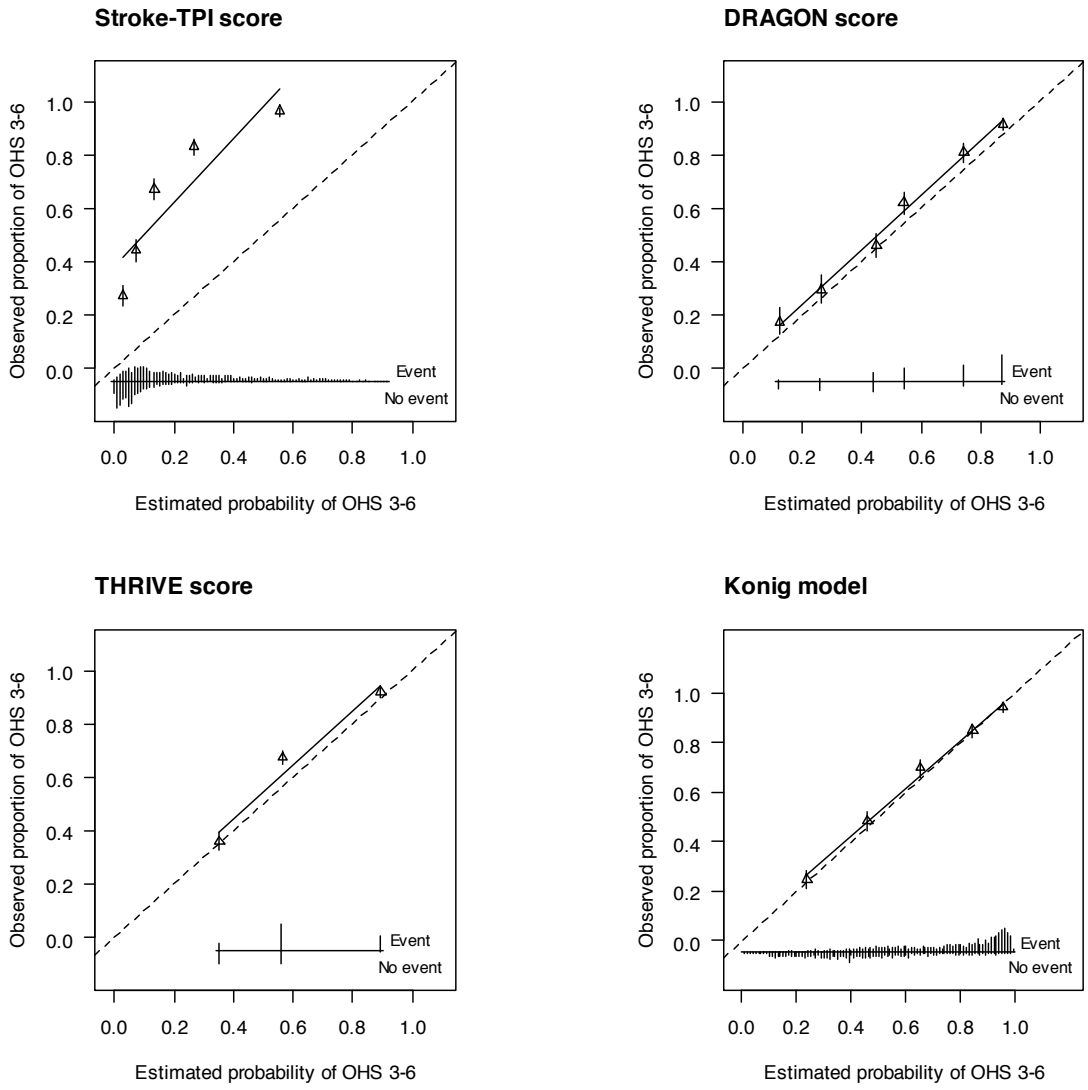


Figure 7-2 Calibration graphs for risk of poor functional outcome (OHS≥3). Each plot shows the predicted risk groups identified either by the model or else grouped as quintiles versus the observed frequencies.

7.3.4.3 Discrimination

Previously developed models to predict SICH or poor functional outcome discriminated modestly between those with and those without an SICH with AUROCC values ranging from 0.56 to 0.65. Models for poor functional outcome post rtPA achieved moderate discrimination with AUROCC values ranging from 0.66 to 0.88 (Table 7-6, Figure 7-3 and Figure 7-4).

The AUROCCs of all models were similar (P-value ≥ 0.05) with the exception of the SPAN-100 score which achieved significantly lower discrimination with SICH (P-value < 0.05) and significantly lower discrimination with poor functional outcome (P-value < 0.001) (see section 7.7 Appendix A on page 236 for the non-parametric Delong tests).

Table 7-6 Discrimination of models to predict intracranial haemorrhage and poor functional outcome after rtPA in IST-3 dataset.

	Seven day SICH post rtPA			Six month poor functional outcome post rtPA		
	n/N	AUROCC (95%CI)	R^2 (%)	n/N	AUROCC (95%CI)	R^2 (%)
SICH models						
HAT	87/1365	0.62 (0.56 to 0.68)	3.23	856/1365	0.71 (0.68 to 0.73)	19.66
SEDAN	87/1365	0.62 (0.56 to 0.69)	3.48	856/1365	0.74 (0.71 to 0.76)	21.97
SITS	85/1357	0.63 (0.58 to 0.69)	3.84	851/1357	0.66 (0.63 to 0.69)	8.87
GRASPS	87/1365	0.63 (0.57 to 0.68)	2.90	856/1365	0.77 (0.74 to 0.79)	26.40
SPAN-100	102/1507	0.56 (0.52 to 0.61)	1.26	957/1507	0.66 (0.64 to 0.68)	17.03
Poor functional outcome models						
Stroke-TPI	87/1365	0.64 (0.58 to 0.69)	3.43	856/1365	0.80 (0.78 to 0.83)	33.95
DRAGON	87/1363	0.65 (0.59 to 0.70)	4.25	855/1363	0.78 (0.76 to 0.81)	29.26
THRIVE	101/1504	0.60 (0.55 to 0.66)	1.71	955/1504	0.76 (0.74 to 0.79)	20.60
König	102/1507	0.63 (0.58 to 0.68)	3.04	957/1507	0.80 (0.77 to 0.82)	32.33

Note AUROCC for SPAN-100 is estimated from logistic regression with the dichotomy SPAN-100 ≥ 100. The R^2 provided here is Nagelkerke's R^2

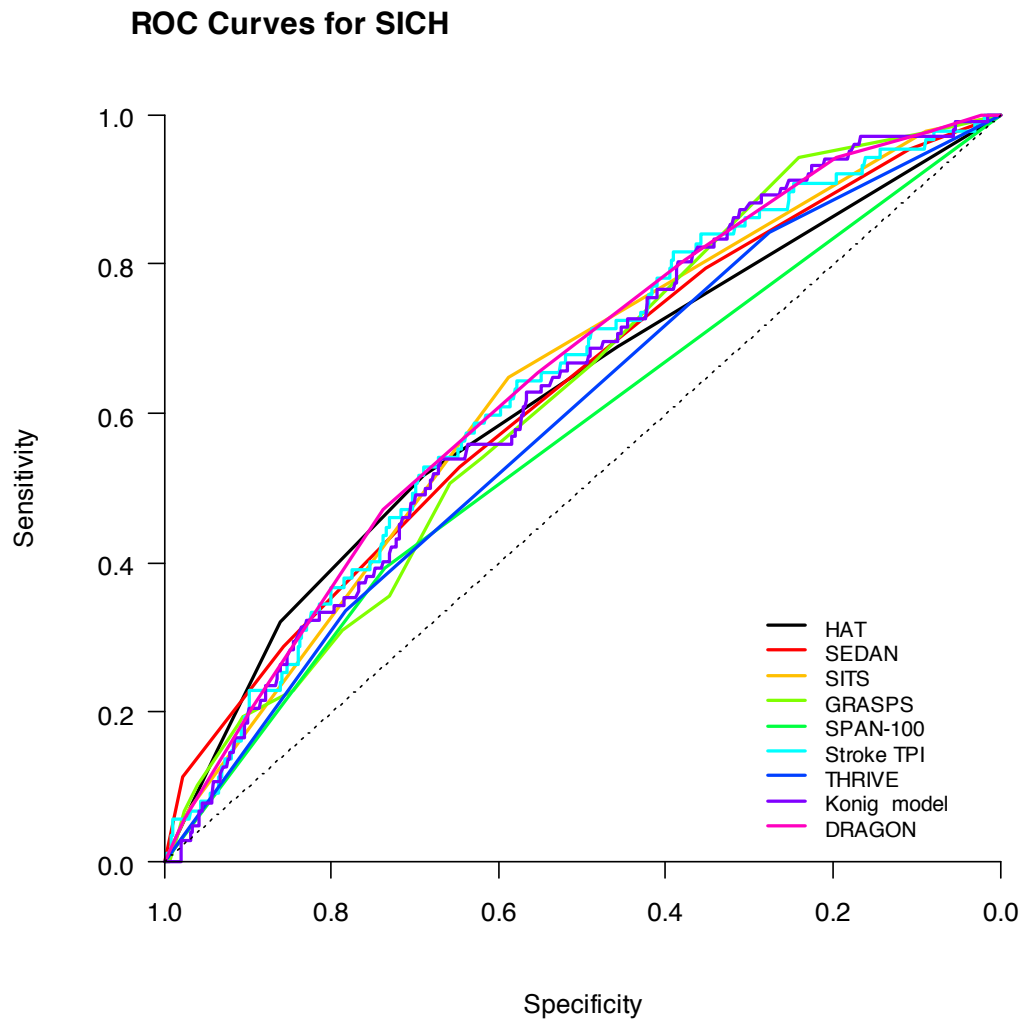


Figure 7-3 Receiver Operating Characteristic (ROC) curves for SICH post rtPA.

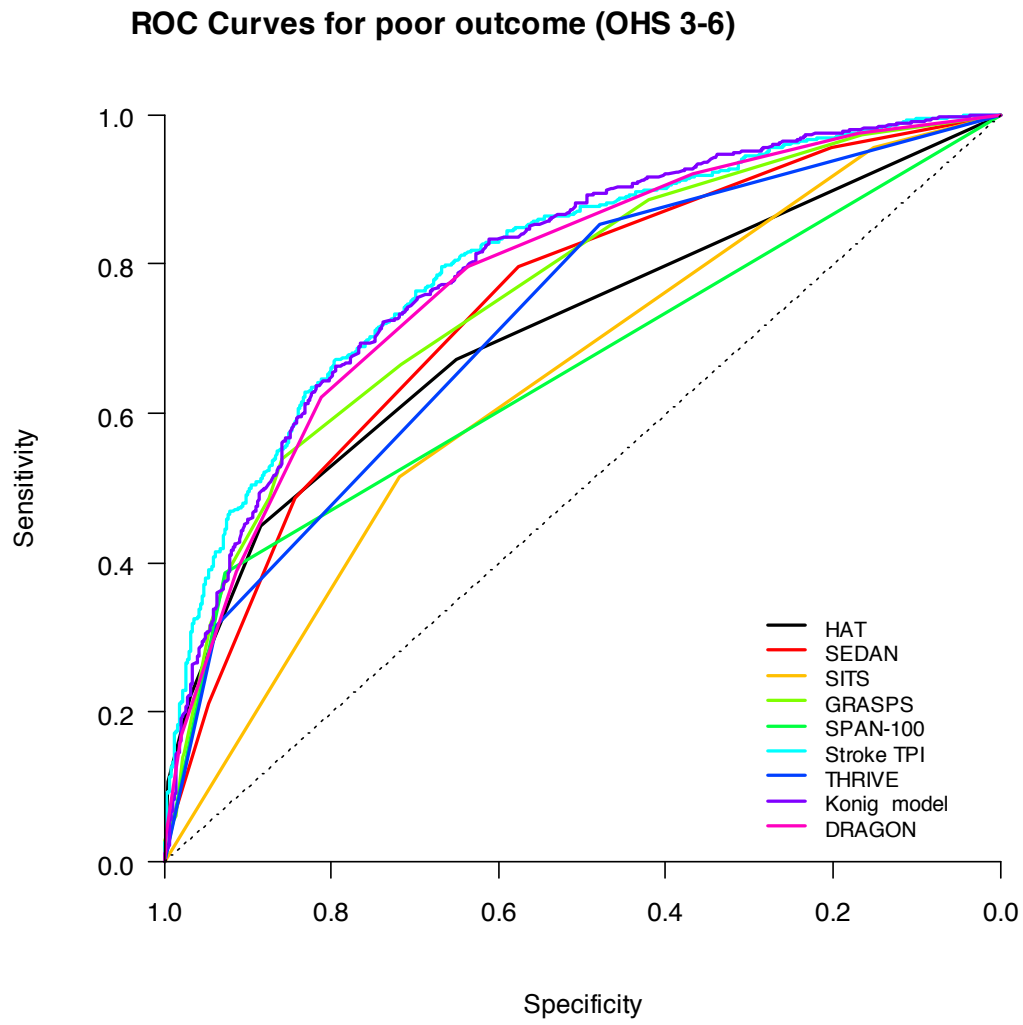


Figure 7-4 Receiver Operating Characteristic (ROC) curves for poor functional outcome (OHS 3-6) post rtPA.

7.4 Model development

The development of a novel prediction model in the IST-3 trial data is of interest as it is often argued that improvements in methodology could have an impact on performance (Royston et al., 2009). A complete case analysis is considered first with an exploration of missing data to follow.

7.4.1 Predictor selection and model assumptions

Predictor selection was defined *a priori* using a systematic review of risk factors associated with SICH post rtPA in acute ischaemic stroke patients (Whiteley et al., 2012). This approach avoids the harmful effects of data dependent selection such as those associated with stepwise selection (see chapter 2 for a more detailed discussion). In brief, data-dependent selection methods have a tendency to over-fit *model* to *data*. A simple way of circumventing this is by pre-specifying a set of clinically important variables *a priori* to include regardless of their achieved statistical significance. Novel methods for estimating parameters in small studies, for example shrinkage of effect estimates using penalized maximum likelihood estimation exist and will be discussed later in this chapter (Steyerberg et al., 2001).

The ratio of events per degree of freedom spent (the 10EPV rule, see Chapter 2) suggests that with 86 post rtPA SICH events observed, no more than nine parameters should be estimated under these data. This was used as an upper threshold placing a necessary limit on the number of parameters that could be reasonably estimated without compromising the model with over-fit estimates. The predictors included in the multivariable binary logistic regression were: age (per decade), NIHSS, glucose (mg/dl), prior hypertension, atrial fibrillation, prior use of anti-platelets, diabetes, leukoaraiosis and visible infarction on CT. These variables were selected from the systematic review by Whiteley *et al*. as variables associated with an increase in the risk of post rtPA SICH evidenced in multiple studies of patients treated with rtPA (Whiteley et al., 2012). Selection was conditional upon the availability of covariates recorded in the IST-3 dataset as well as the associated EPV. All continuous predictors entered the model in their original continuous form – categorising these variables would discard useful information (Altman and Royston, 2006). The

functional form of the continuous predictors was assessed using restricted cubic splines (RCS), i.e., for age, NIHSS, and glucose, varying the number of knots over 3, 4 and 5. The RCS function provides a versatile assessment of the relationship between outcome and predictor and was used to test how plausible a non-linear relationship was beyond that of the simple linear fit. The additivity assumption was assessed by comparing nested models with and without two-way interactions between the continuous variables and the binary variables: diabetes, visible infarct on CT, anti-platelets, leukoaraiosis, hypertension, and atrial fibrillation. This formally tests the hypothesis that each effect has a constant impact on the log odds of the risk from post rtPA SICH. Although the model χ^2 was larger under the more complicated RCS fit (Figure 7-5), after penalizing for the additional parameters estimated this improvement was either removed or was more conservative (Table 7-7).

There was insufficient evidence to support the inclusion of any two way interactions or any non-linear relationships with the log odds of SICH.

The final model is provided in Table 7-8. Higher NIHS scores were associated with a statistically significant increase in the odds of SICH (1.06 with 95%CI: 1.03 to 1.09, P-value = 0.0005), as was prior anti-platelet use (2.15 with 95%CI: 1.32 to 3.50, P-value = 0.0022). History of atrial fibrillation was associated with a significant decrease in the odds of SICH (0.58 with 95%CI: 0.34 to 0.97, P-value = 0.0392).

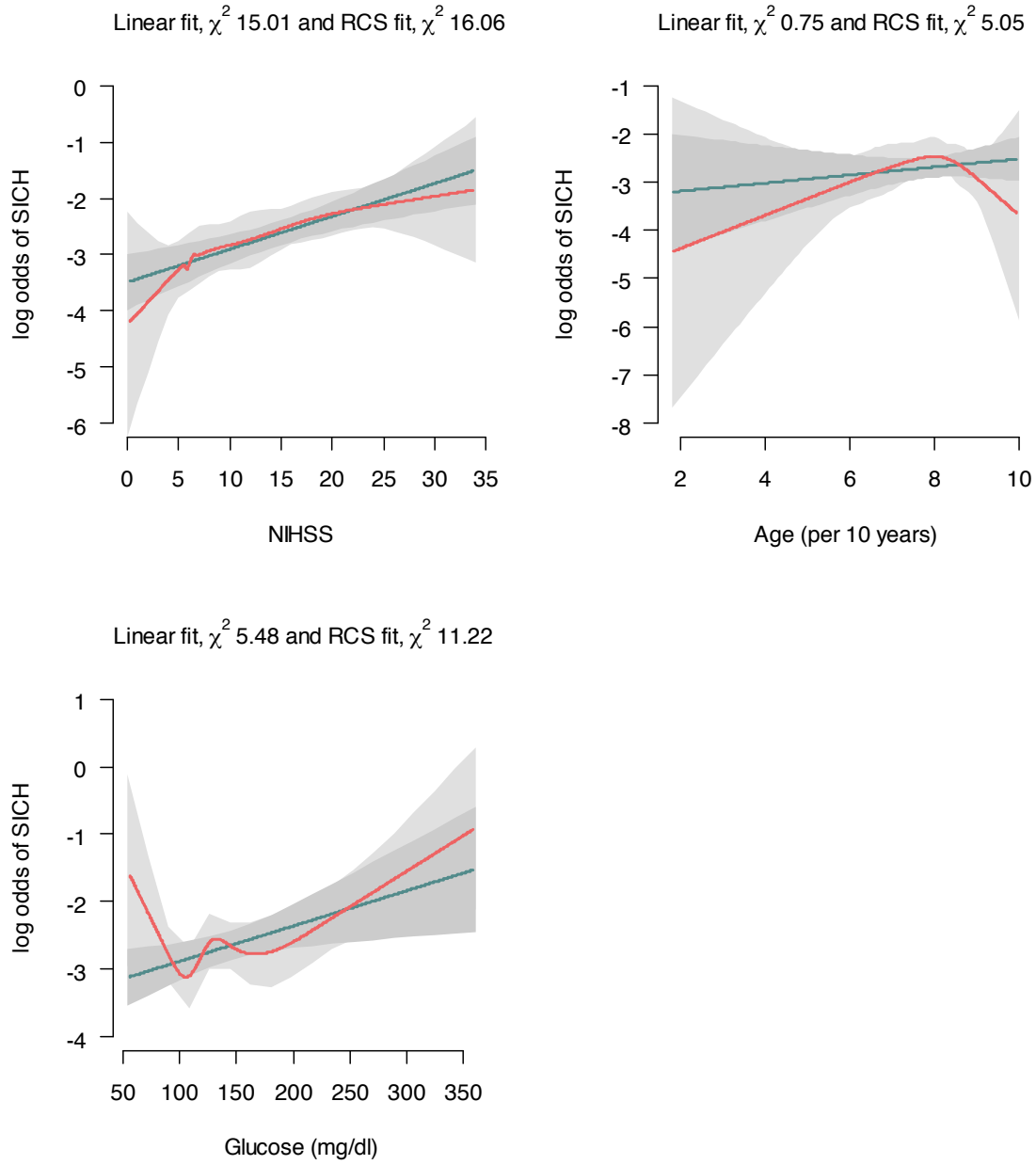


Figure 7-5 Transformations of NIHSS, age per decade and glucose in univariate analysis comparing a simple linear fit to a flexible restricted cubic spline with 5 knots.

Table 7-7 Testing model assumptions for seven day SICH post rtPA model in the IST-3 dataset.

	Seven day SICH post rtPA		
	Age (per decade)	**NIHSS**	**Glucose (mg/dl)**
Additivity			
Diabetes	-1.04 (0.3073)	0.40 (0.5291)	2.72 (0.0992)
Visible infarct on CT	-1.95 (0.1623)	-1.42 (0.2341)	2.23 (0.1351)
Antiplatelet	0.92 (0.3379	-1.65 (0.1996)	-1.93 (0.1649)
Leukoaraiosis	-1.12 (0.2905)	1.41 (0.2351)	2.52 (0.1122)
Hypertension	-1.87 (0.1715)	-1.68 (0.1954)	-1.98 (0.1598)
Atrial Fibrillation	-1.29 (0.2559)	-2.00 (0.1576)	-0.03 (0.8522)
Linearity			
Restricted cubic spline			
Three knots	2.68 (0.1017)	-0.65 (0.4204)	-0.57 (0.4488)
Four knots	0.83 (0.6602)	-2.42 (0.2976)	-1.27 (0.5297)
Five knots	-1.01 (0.7990)	-4.32 (0.2286)	0.33 (0.9553)
Square term	2.06 (0.0439)	-0.33 (0.1962)	-1.15 (0.3570)

NOTE: The difference in AIC units contrasting a complex fit to a simple fit is provided alongside the P-value for the associated LR test. Positive AIC on the chi-squared scale represent an improved fit

Table 7-8 Multivariable logistic regression model for seven day risk of SICH post rtPA based on complete case data (86/1361).

Variable	Coefficient (SE)	Wald	OR (95% CI)	P-value
Intercept	-4.461 (0.929)	-4.80	-	<0.0001
Age (per decade)	-0.013 (0.116)	-0.11	0.99 (0.79 to 1.24)	0.9102
NIHSS	0.057 (0.017)	3.47	1.06 (1.03 to 1.09)	0.0005
Glucose (mg/dl)	0.005 (0.002)	1.92	1.00 (1.00 to 1.01)	0.0553
Prior Hypertension	-0.121 (0.251)	-0.48	0.89 (0.54 to 1.45)	0.6297
Atrial Fibrillation	-0.551 (0.267)	-2.06	0.58 (0.34 to 0.97)	0.0392
Visible infarct on CT	0.228 (0.239)	0.95	1.26 (0.79 to 2.01)	0.3404
Anti-platelet	0.764 (0.249)	3.06	2.15 (1.32 to 3.50)	0.0022
Diabetes	0.089 (0.352)	0.25	1.09 (0.55 to 2.18)	0.8005
Leukoaraiosis	0.236 (0.248)	0.95	1.27 (0.78 to 2.06)	0.3415

7.4.2 Missing data

The handling of missing data in this analysis was primarily a sensitivity analysis. If the results of a complete-case analysis did not differ qualitatively from that of an analysis based on imputation then the complete-case data would be favored for its simplicity. Indeed under a complete-case analysis only 10% of those treated with rtPA were excluded (Table 7-8). For a clinical audience a complete-case model would be easier to accept. The extent of missingness was small amongst those risk factors known to be associated with post rtPA SICH which, with the exception of blood glucose, were either completely observed or had fewer than 1% missing (Table 7-9). Blood glucose was not recorded for the first 283 patients entering the trial and can therefore be assumed to be Missing At Random (MAR).

Of those missing variables listed in Table 7-9, 1945 (64%) patients had completely observed data, 403 (13%) had one missing observation, 406 (13%) had two missing observations, 139 (5%) had three missing observations whilst fewer had multiple missing values. The maximum number of missing observations per patient was eight, of which there was only one. A cluster plot was used to visualize joint missingness (Figure 7-6). Impairments caused by the initial stroke were often jointly missing and along with blood glucose and prior use of warfarin and/or heparin, contributed the majority of missingness. Ascertainment of patient data was good. The imputation model used all of those variables listed in Table 7-9. The Missing At Random (MAR) mechanism could be assessed in part using logistic regression. Indicator variables for missing values for each variable with missing data where used to establish to what extent missingness could be described by the observed values amongst the other predictors. Explained variation ranged from 53% for weakness in leg after stroke to 4% for other deficits suggesting that values were therefore at least partly missing at random and that a multiple imputation approach could reduce the risk of bias. The model based on complete-case data was compared to the model averaged over 50 imputations. This made little material difference to the magnitude or direction of the estimates (beta coefficients shown in Figure 7-7) or the variation explained (Nagelkerke's R^2 in MI fit 7.15%, and CC fit 6.94%). For ease of interpretation the complete-case model was taken forward.

Table 7-9 Baseline characteristics of IST-3 patients. The median and IQR is provided for continuous variables and the frequency and percentage for categorical variables.

	rtPA (N=1515)		Control (N=1520)	
Characteristic	**Measure**	**Missing**	**Measure**	**Missing**
Prediction model variables				
Age, years	81 (72-86)	0 (0)	81 (71-86)	0 (0)
NIHSS	11 (6-18)	0 (0)	11 (6-17)	0 (0)
Blood glucose (mg/dl)	126 (108-144)	142 (9)	126 (108-144)	141 (9)
Prior hypertension	975 (64)	2 (<1)	979 (64)	6 (<1)
Atrial Fibrillation	473 (31)	0 (0)	441 (29)	0 (0)
Visible infarct on CT	628 (41)	0 (0)	604 (40)	0 (0)
Taking antiplatelets	775 (51)	4 (<1)	787 (52)	7 (<1)
Diabetes	184 (12)	2 (<1)	204 (13)	2 (<1)
Presence of leukoaraiosis	765 (50)	8 (1)	782 (51)	10 (1)
Additional variables				
Gender, male	733 (48)	0 (0)	732 (48)	0 (0)
Weight (kg)	70 (62-80)	0 (0)	70 (60-80)	0 (0)
Systolic BP (mmHg)	156 (140-170)	0 (0)	155 (140-170)	0 (0)
Diastolic BP (mmHg)	80 (71-91)	12 (1)	80 (72-90)	7 (<1)
Delay from randomisation (hours)	3.8 (2.9-4.8)	0 (0)	3.9 (2.9-4.8)	0 (0)
Prior stroke or TIA	354 (23)	3 (<1)	345 (23)	2 (<1)
Taking warfarin or heparin	50 (3)	149 (10)	44 (3)	149 (10)
Imaging variables				
Hyperdense artery	376 (25)	8 (1)	359 (24)	10 (1)
Deficits caused by stroke				
Weakness in face	1255 (83)	12 (1)	1267 (83)	15 (1)
Weakness in arm	1300 (86)	4 (<1)	1322 (87)	6 (<1)
Weakness in leg	1189 (78)	10 (1)	1208 (79)	13 (1)
Presence of dysphasia	775 (51)	39 (3)	809 (53)	26 (2)
Presence of hemianopia	588 (39)	226 (15)	576 (38)	251 (17)
Visuospatial disorder	550 (36)	282 (19)	532 (35)	329 (22)
Brainstem signs	106 (7)	105 (7)	132 (9)	117 (8)
Other deficits	173 (11)	96 (6)	176 (12)	90 (6)

ABBREVIATIONS: NIHSS - National Institutes of Health Stroke Scale; BP - Blood Pressure; mmHg - millimeter of mercury; mg/dl – milligram per deciliter; kg – Kilogram; TIA – Transient Ischaemic Attack

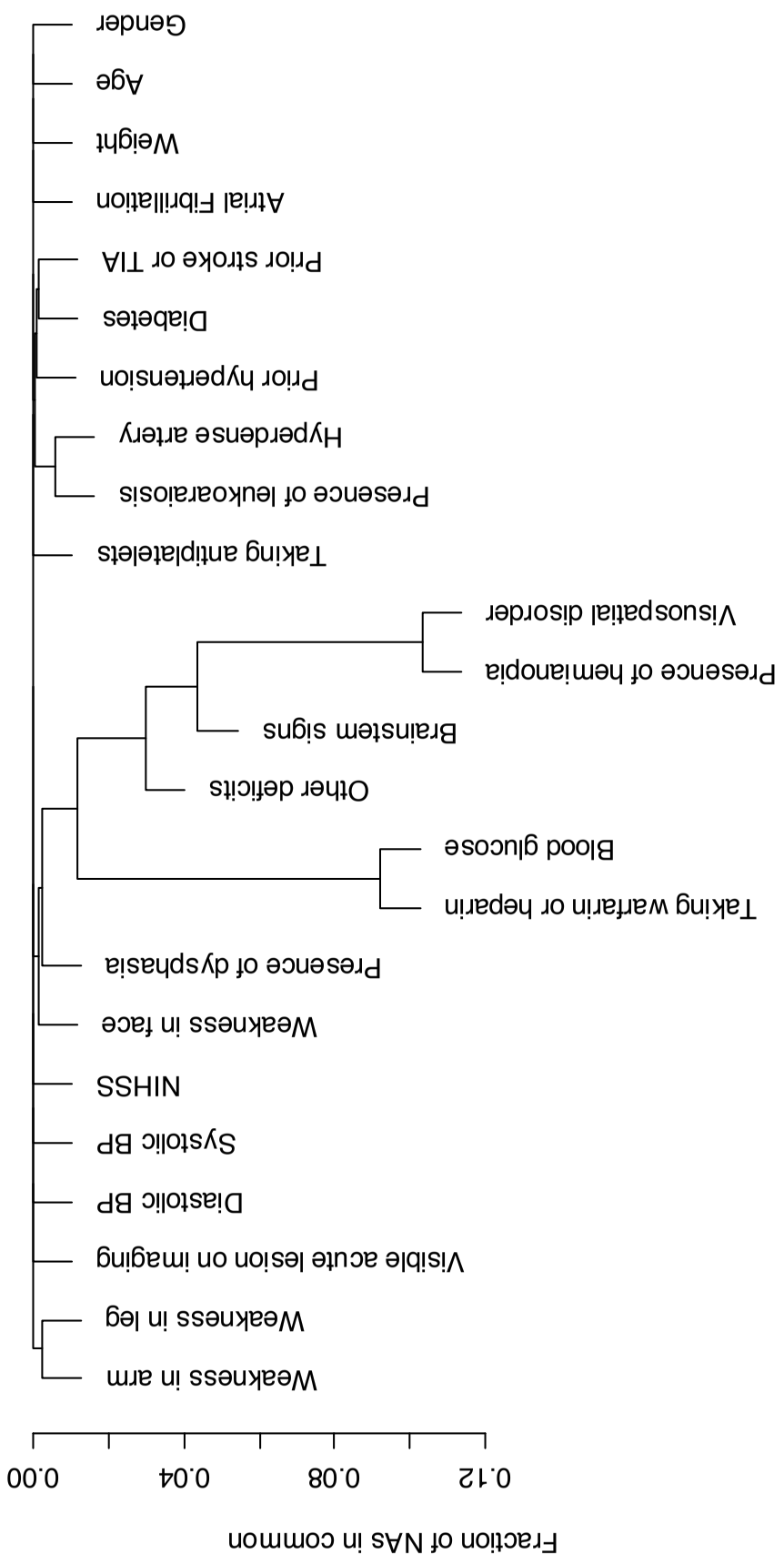


Figure 7-6 Combinations of missing values in IST-3; a hierarchical cluster analysis of combined missingness.

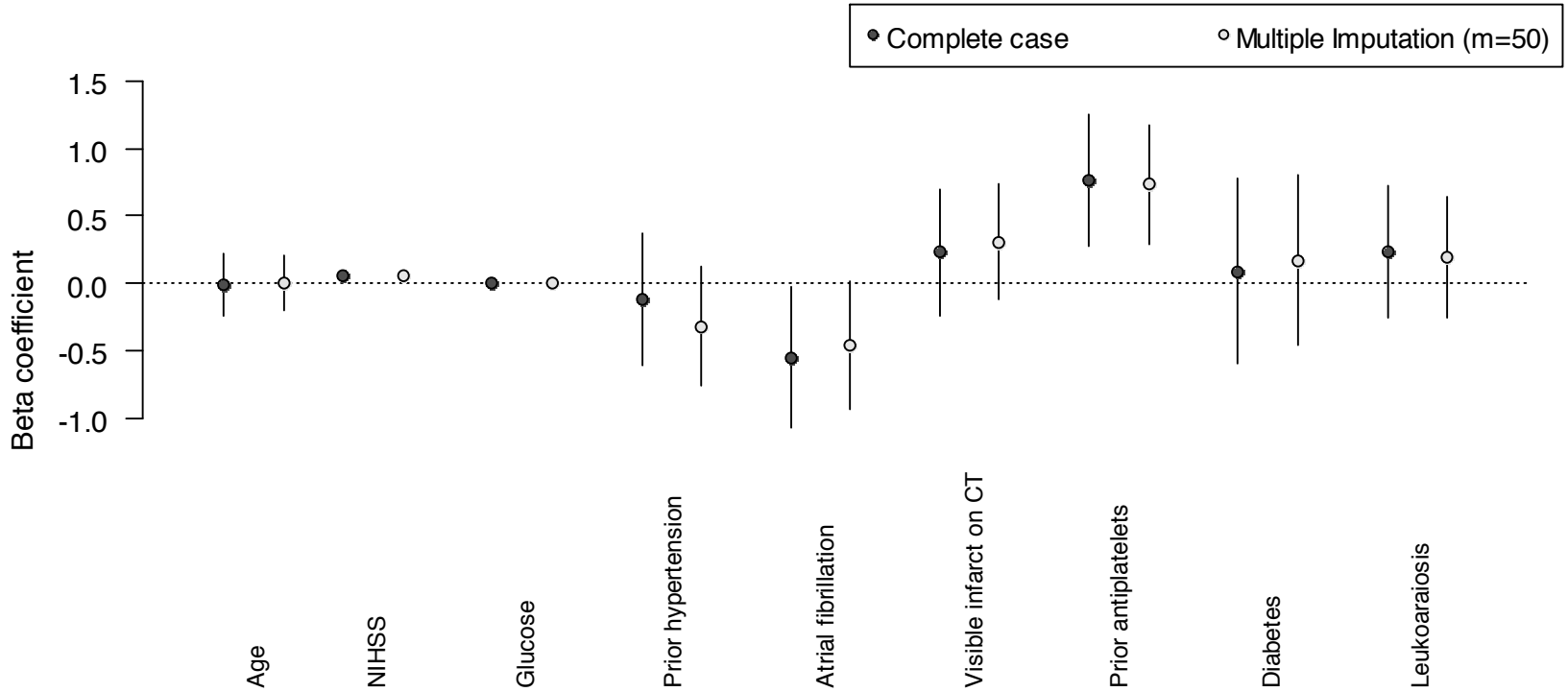


Figure 7-7 Comparison of two models in the IST-3 data for the prediction of post rtPA SICH: complete-case (dark grey) versus imputation across 50 imputed sets (light grey).

7.4.3 The added predictive value of brain imaging variables

The use of novel brain imaging techniques may improve the prediction of SICH (Whiteley et al., 2012). The IST-3 protocol stipulated that prior to randomisation all patients must have received a brain scan so as to exclude any haemorrhagic strokes. Additional scans were taken between 24 and 48 hours from randomisation (Sandercock et al., 2011). The preference was for a computed tomography (CT) scan of the entire brain, though if not possible then magnetic resonance (MR) imaging could be used provided the radiological support was available to interpret the obtained scans.

A sub-study within IST-3 was devised to evaluate novel techniques or scores for brain imaging. The aim was to assess how *perfusion* and *angiographic* imaging using CT or MRI may be used in selecting those patients most likely to benefit from treatment with rtPA (Wardlaw et al., 2012b). Specifically, the identification of dense arteries or substantial changes in white matter may help avoid treating those likely to suffer harm from rtPA. Cerebral perfusion describes the flow of blood throughout the vascular network in the brain; a CT or MRI has the potential to determine the amount of tissue that remains at risk (Petrella and Provenzale, 2000). The Alberta Stroke Program Early CT (Computed Tomography) score (ASPECTS) is a ten-point scoring system which aims to improve the reliability of ischaemic change in the middle cerebral artery (MCA) as identified through CT (Pexman et al., 2001). Here a lower score is associated with a worse prognosis. Previous studies have suggested an association with an increased risk from SICH (Puetz et al., 2009).

The total ASPECTS score was associated with SICH post rtPA (OR 0.90 with 95% CI: 0.83 to 0.97 and P-value = 0.0053) at the univariate level in the IST-3 data. When this was adjusted for those predictors used in the multivariable model (Table 7-8) the association was no longer significant (OR 0.96 with 95% CI: 0.85 to 1.07 and P-value = 0.4524). The AUROCC statistic increased only slightly by 0.003 (P-value = 0.4343) and the Nagelkerkes R^2 by 0.11%. Similar findings were made for the presence of hyperdense arteries with regards to: association (univariate OR 1.86 (95% CI: 1.17 to 2.96) with P-value = 0.0092; and multivariable OR 1.43 (95% CI:

0.83 to 2.47) with P-value = 0.2014); discrimination (difference in AUROCC of 0.006, P-value = 0.3720); and the variation explained (difference in Nagelkerkes R^2 of 0.31%).

Reclassification metrics were used to examine the utility of scan measurements to classify patients (Pencina et al., 2008). These were introduced in Chapter 2. Strata of risk were defined and reclassification tables produced. These were supplemented by the categorised NRI. Low (<3%), medium (3 to 8%), and high (>8%) risk categories of SICH risk were defined by surveying the thresholds reported in previous studies and using the means of these thresholds. Reclassification plots were produced and supplemented by the continuous NRI thus exploring the case where no thresholds are defined (Pencina et al., 2011).

For total ASPECTS (Table 7-10), the net-reclassification amongst events (N = 86) was +2.33% but was not statistically different from zero (P-value = 0.6543) with only three patients (3/86) correctly reclassified at greater risk and one (1/86) at lower risk. Amongst non-events (N = 1275) the net-reclassification was -0.31% which was also not statistically different from zero (P-value = 0.5862), here, 25 patients (25/1275) incorrectly moved up to a greater risk category whilst 29 (29/1275) correctly moved down. The total NRI was 0.0264 (95% CI: -0.0206 to 0.0733 with P-value = 0.2706).

For visible hyperdense arteries (Table 7-11), the net-reclassification amongst events was +1.16% which was not statistically different from zero (P-value = 0.3145) with three patients (3/86) correctly reclassified at greater risk and two (2/86) incorrectly classified as lower risk. Amongst non-events the net-reclassification was 0% since 55 patients (55/1275) incorrectly moved up to a greater risk category whilst 55 (55/1275) correctly moved down. The total NRI was 0.0116 (95% CI: -0.0418 to 0.0651 with P-value = 0.6698).

Table 7-10 Reclassification table for the predicted probabilities with and without the imaging variable, total ASPECTS. NRI was 0.0264 (95% CI: -0.0206 to 0.0733, P-value= 0.2706).

Model without ASPECTS	Model with ASPECTS			
Frequency (Row %)	< 3 %	3 to 8 %	> 8 %	Total
Participants who experience a SICH event				
< 3 %	5 (71)	**2 (29)**	**0 (0)**	7 (8)
3 to 8 %	0 (0)	38 (97)	**1 (3)**	39 (45)
> 8 %	0 (0)	1 (2)	39 (98)	40 (47)
Total	5 (6)	41 (48)	40 (47)	86
Participants who did not experience a SICH event				
< 3 %	269 (95)	15 (5)	0 (0)	284 (22)
3 to 8 %	**10 (1)**	677 (97)	10 (1)	697 (55)
> 8 %	**0 (0)**	**19 (6)**	275 (94)	294 (23)
Total	279 (22)	711 (56)	285 (22)	1275

Note: dark grey shaded area indicates movement in the correct direction whilst light shaded area indicates movement in the wrong direction

Table 7-11 Reclassification table for the predicted probabilities with and without the visible hyperdense arteries. The NRI was 0.0116 (95% CI: -0.0418 to 0.0651, P-value= 0.6698).

Model without hyperdense arteries	Model with hyperdense arteries			
Frequency (Row %)	< 3 %	3 to 8 %	> 8 %	Total
Participants who experience a SICH event				
< 3 %	7 (100)	**0 (0)**	**0 (0)**	7 (8)
3 to 8 %	0 (0)	36 (92)	**3 (8)**	39 (45)
> 8 %	0 (0)	2 (5)	38 (95)	40 (47)
Total	7 (8)	38 (44)	41 (48)	86
Participants who did not experience a SICH event				
< 3 %	260 (92)	24 (8)	0 (0)	284 (22)
3 to 8 %	**20 (3)**	646 (93)	31 (4)	697 (55)
> 8 %	**0 (0)**	**35 (12)**	259 (12)	294 (23)
Total	280 (22)	705 (55)	290 (23)	1275

Note: shaded areas defined as above in Table 7-10

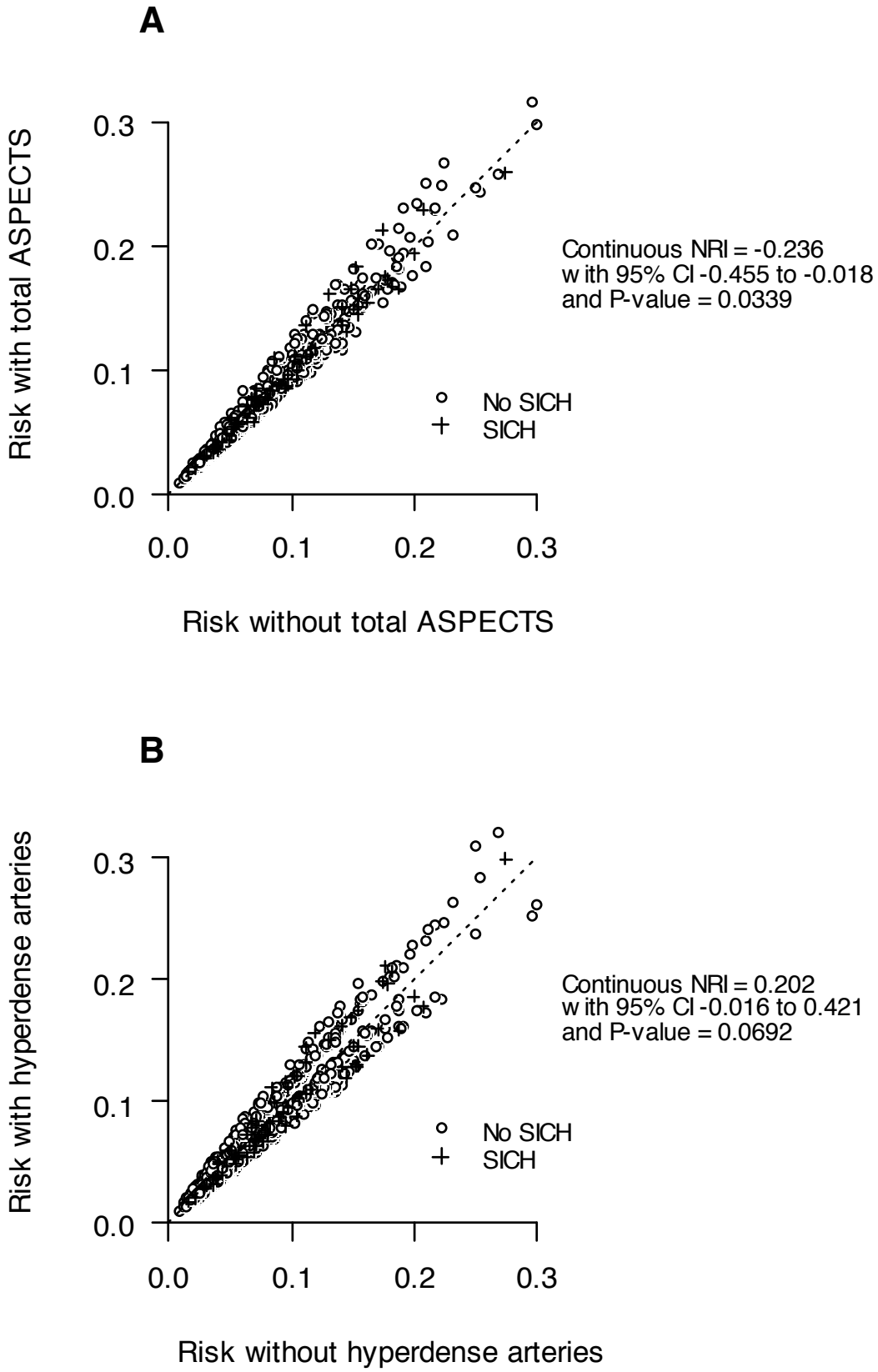


Figure 7-8 Reclassification plots: (A) Total ASPECTS; and (B) hyperdense arteries.

A more flexible assessment of reclassification is provided by the continuous NRI which does not impose any pre-specified, and potentially arbitrary, risk categories, but instead quantifies any upward or downward movement in risk (Pencina et al., 2011). This is demonstrated in Figure 7-8.

For total ASPECTS the continuous net-reclassification for SICH was -11.60% (95% CI: -3.28% to +9.51%) with a P-value of 0.2810 and for no SICH was -12.00% (95% CI: -17.50% to -6.51%) with a P-value <0.0001. For the presence of hyperdense arteries the continuous net-reclassification for SICH was -20.90% (95% CI: -42.07% to +0.21%) with a P-value of 0.0523 and for no SICH was +41.20% (95% CI: +35.69% to +46.67%) with a P-value <0.0001.

A detailed discussion was provided in Chapter 2 outlining the many concerns that are associated with these NRI statistics. In particular the continuous NRI which attributes equal weight to any movement in classification. It is likely that this will give undue importance to small shifts amongst those at low risk resulting in a larger NRI value which is void of any meaningful interpretation.

If an added predictor is not significant when assessed by a simple LR test then ultimately any exploration of reclassification is futile (Cook, 2007). Using the LR test it was found that the inclusion of total ASPECTS did not correspond to a significant improvement in the model chi-squared (a gain in model χ^2 of 0.56 with a P-value of 0.4557). Similarly with the presence of hyperdense arteries there was little to be gained by including this predictor (a gain in model χ^2 of 1.61 with a P-value of 0.2047).

7.4.4 Shrinkage of regression coefficients

Maximum likelihood (*ML*) estimation tends to overstate the evidence provided by small data sets producing an optimal fit that is unlikely to generalise well. By shrinking the estimated coefficients the model may have better application post development (Steyerberg et al., 2001). A heuristic estimate of shrinkage can be used to estimate just how much of the model fit is noise (Van Houwelingen and Le Cessie, 1990):

$$\hat{\gamma} = \frac{model\ \chi^2 - p}{model\ \chi^2} \tag{7.1}$$

For the SICH model given in Table 7-8 the model χ^2 was 35.97 with nine degrees of freedom, it is expected that 25% of this fit was noise. This model is unlikely to perform well in new data. The Penalised Maximum Likelihood (*PML*) estimation method implements an active shrinkage of regression coefficients during estimation with a dynamic penalty (Steyerberg, 2009).

$$PML = \hat{l} - \frac{1}{2}\lambda\beta^T P\beta \tag{7.2}$$

Where λ is the penalty factor, β is the vector of estimated predictor effects (which excludes the intercept), β^T is the transpose of β and P is a non-negative penalty matrix which, for k estimated parameters, can be set as the diagonal matrix of the k variance estimates, i.e., $diag(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \cdots, \hat{\sigma}_k^2)$. It is noted that if the penalty factor is set equal to zero then the standard log likelihood is used. Also, since the expression $0.5\ \lambda\beta^T P\ \beta$ is non-negative, the *PLM* will always be less than $\hat{l}$ provided that $\lambda>0$. Choosing an optimal value for λ is undertaken by maximising a modified version of the Akaike's Information Criterion (AICc). The AICc is used because it corrects for small sample sizes by applying a heavier penalty for additional parameters.

$$AICc = AIC + \frac{2(df_{effective} + 1)}{n - df_{effective} - 1} \tag{7.3}$$

Where $df_{effective}$ is the effective degrees of freedom and n is the size of the sample. The effective degrees of freedom describe the reduction in variance achieved when moving from standard *ML* estimation to *PML* estimation and is given by,

$$df_{effective} = tr[I(\beta)Cov(\beta)].$$

Where $I(\beta)$ is the information matrix, or *Fisher's information*, calculated without penalty (the expected value of the second derivative of the log likelihood) and $Cov(\beta)$ as the inverse of $I(\beta)$ calculated with penalty (note: in linear algebra $tr[\cdot]$ denotes the trace of a matrix which is simply the sum of the on-diagonal values). By varying λ across a fine grid of values an optimal value which maximises the AICc can be obtained (shrinkage of the model presented in Table 7-8 yields a maximised AICc of 20.99 with a penalty of 19.35).

The model presented in Table 7-12 will be used as the novel model and will be referred to as the *IST-3 model*.

Table 7-12 *PML* multivariable logistic regression model for 7 day risk of SICH post rtPA based on complete case data (86/1361).

Variable	Coefficient (SE)	Wald	OR (95% CI)	P-value	Penalty
Intercept	-4.203 (0.796)	-5.28	-	<0.0001	0.00
Age (per decade)	0.005 (0.095)	0.05	1.00 (0.83 to 1.21)	0.9617	5.26
NIHSS	0.045 (0.014)	3.16	1.05 (1.02 to 1.08)	0.0016	31.29
Glucose (mg/dl)	0.004 (0.002)	1.89	1.00 (1.00 to 1.01)	0.0584	192.54
Prior Hypertension	-0.041 (0.196)	-0.21	0.96 (0.65 to 1.41)	0.8337	3.11
Atrial Fibrillation	-0.304 (0.120)	-1.52	0.74 (0.50 to 1.09)	0.1290	3.11
Visible infarct on CT	0.186 (0.190)	0.98	1.20 (0.83 to 1.75)	0.3291	3.11
Anti-platelet	0.482 (0.191)	2.52	1.62 (1.11 to 2.36)	0.0116	3.11
Diabetes	0.064 (0.236)	0.27	1.07 (0.67 to 1.69)	0.7849	3.11
Leukoaraiosis	0.146 (0.193)	0.75	1.16 (0.79 to 1.69)	0.4506	3.11

.

7.4.5 Model performance: internal evaluation

With only the IST-3 data available to evaluate performance an internal evaluation using repeated bootstrap sampling (150 replicates) was followed. This expressed the likely performance of the *ML* SICH post rtPA fit and the *PML* fit. The estimates of performance obtained through internal evaluation correct for the optimism expressed when evaluating the model in the original data (apparent performance). The AUROCC for the *PML* fit suggested moderate discrimination (Table 7-13) which was slightly more conservative (0.65) than the apparent performance (0.68). Note that as of yet there is no agreed method for obtaining confidence intervals for estimates obtained from internal evaluation in this way.

Table 7-13 Internal evaluation of IST-3 model (Table 7-12) in 150 bootstrap replicates.

	Apparent	Training	Test	Optimism	Optimism corrected
Standard *ML* model					
Nagelkerkes R^2 (%)	6.940	9.283	5.674	3.609	3.331
AUROCC	0.688	0.712	0.670	0.042	0.646
Slope	1.000	1.000	0.787	0.213	0.787
Intercept	0.000	0.000	-0.511	0.511	-0.511
***PML* model**					
Nagelkerkes R^2 (%)	5.175	6.218	5.781	0.437	4.738
AUROCC	0.685	0.699	0.670	0.029	0.646
Slope	1.000	1.000	1.148	-0.148	1.148
Intercept	0.000	0.000	0.400	-0.400	0.400

7.5 A stratified treatment approach for rtPA

Having assessed the performance of pre-existing models and explored the potential for an increase in performance where more robust methods have been used in model development, the remainder of this chapter will explore the hypothesis that treatment effect varies according to predicted risk.

7.5.1 Recalibration using a simple updating procedure

Predictions were made for each patient recruited to the IST-3 using the prediction scores and models discussed in sections 7.3.2 and 7.3.3. If a full regression model was available with an associated intercept term then a simple updating procedure was used adjusting for the difference in incidence of the outcome between the data used in development and the IST-3 data (Janssen et al., 2009). This form of adjustment relates only to the estimated risk and does not affect the ordering of the patients; therefore, discrimination – a rank based measure – is unaffected. As noted earlier in Chapter 6, a correction factor is calculated from the incidence of the outcome in the new data and the mean predicted risk estimated by the model using equation (6.1).

This correction factor (*cf*) can be interpreted as the log odds ratio of incidence and estimated risk. Adding this *cf* to the original model constitutes as updating the old model to the new local setting. Recalibration was possible for three of the ten models: the Stroke-TPI model; the König model and the IST-3 model (see section 7.8 Appendix B on page 240). Two of which were pre-existing models for the prediction of poor functional outcome and one which was developed on the IST-3 data for the prediction of post rtPA SICH. Each therefore required updating for the incidence of post rtPA SICH and/or poor functional outcome (OHS≥3) specific to the IST-3 patients. Low, medium and high risk strata were defined separately for the predicted risk of SICH and for poor functional outcome by surveying those defined categories used by authors of previous models. The distribution of these strata varied depending on the risk score used (Table 7-14).

Table 7-14 Total number within each model predicted risk strata (N, %) with associated number of poor outcomes (n, %). Note: in some cases risk strata were left undefined as no patients were classified by these models. This was due to a limited number of estimated categories available in some score models.

Models	SICH (n/N)			Poor functional outcome, OHS≥3 (n/N)		
	Low (<3%)	Medium (3 to 8%)	High (>8%)	Low (<35%)	Medium (35 to 56%)	High (>56%)
HAT	581 (33)/1246 (46)	392 (23)/626 (23)	763 (44)/863 (32)	-	-	1736 (100)/2735 (100)
SEDAN	370 (21)/956 (35)	-	1366 (79)/1779 (65)	-	-	1736 (100)/2735 (100)
SITS	832 (48)/1523 (56)	815 (47)/1099 (40)	75 (4)/93 (3)	-	-	1722 (100)/2715 (100)
SPAN-100	-	-	1939 (100)/3017 (100)	-	-	1939 (100)/3017 (100)
GRASPS	206 (12)/649 (24)	1131 (65)/1657 (61)	399 (23)/429 (16)	-	-	1736 (100)/2735 (100)
Stroke-TPI	703 (40)/1574 (58)	496 (29)/603 (22)	537 (31)/558 (20)	282 (22)/903 (33)	389 (30)/630 (23)	1065 (48)/1202 (44)
THRIVE	-	-	1934 (100)/3009 (100)	290 (15)/809 (27)	1039 (54)/1542 (51)	605 (31)/658 (22)
DRAGON	-	-	1734 (100)/2732 (100)	123 (7)/506 (19)	547 (32)/1007 (37)	1064 (61)/1219 (45)
König model	731 (38)/1652 (55)	561 (29)/677 (22)	647 (33)/688 (23)	269 (14)/885 (29)	437 (23)/728 (24)	1233 (64)/1404 (47)
IST-3 model	54 (3)/237 (9)	1143 (66)/1874 (69)	532 (31)/612 (22)	1 (<1)/7 (<1)	415 (24)/948 (35)	1313 (76)/1768 (65)

7.5.2 Absolute risk reduction

For each of the risk groups per model described in Table 7-14 a two by two table cross-classifying patient outcome (i.e., OHS<3 vs. OHS≥3) with treatment received (i.e., rtPA vs. control) was produced. The absolute risk reduction (*ARR*) was then calculated as the difference in the conditional probabilities for poor outcome in the control and in the treatment arms. Positive *ARR* values denote benefit whilst negative values denote harm. An estimate for the variance can be obtained to provide an asymptotic 95% CI.

The following two figures (Figure 7-9 and Figure 7-10) illustrate the absolute estimated effect of rtPA according to strata of predicted risk from SICH and from poor functional outcome amongst the IST-3 patients. The size of each point is related to the proportion of patients categorised by a given model.

In general, the *ARR* of poor functional outcome with rtPA was at its greatest favoring the active treatment amongst those patients with high predicted risk of SICH (Figure 7-9) or high predicted risk of poor functional outcome (Figure 7-10). Under this categorisation there was no indication of significant harm amongst those with a lower predicted risk, i.e., a 95% CI with an upper limit less than zero.

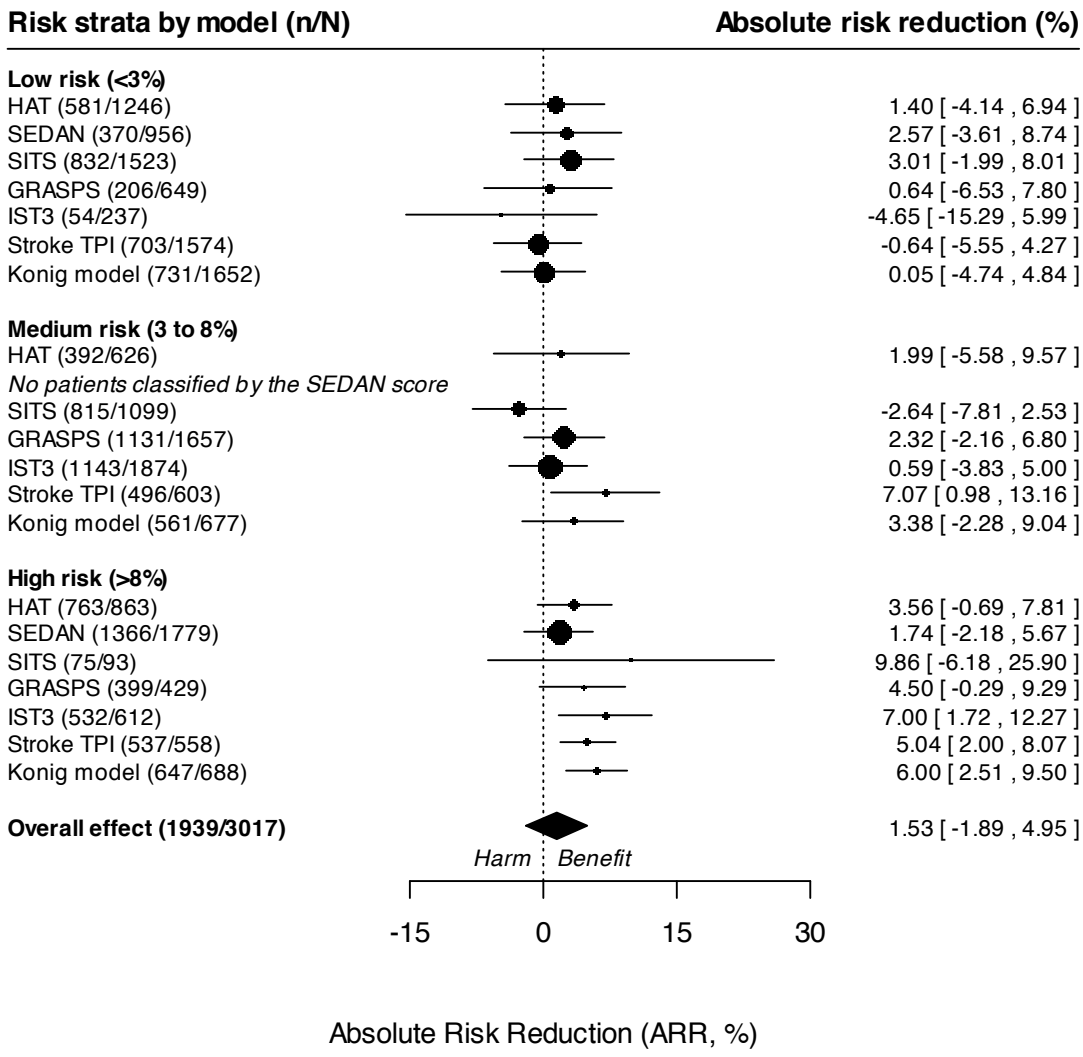


Figure 7-9 Effect of rtPA on six month OHS≥3 stratified by predicted SICH risk. No patients were classified 'medium risk' by the SEDAN score. Point size is relative to the size of the denominator across risk strata.

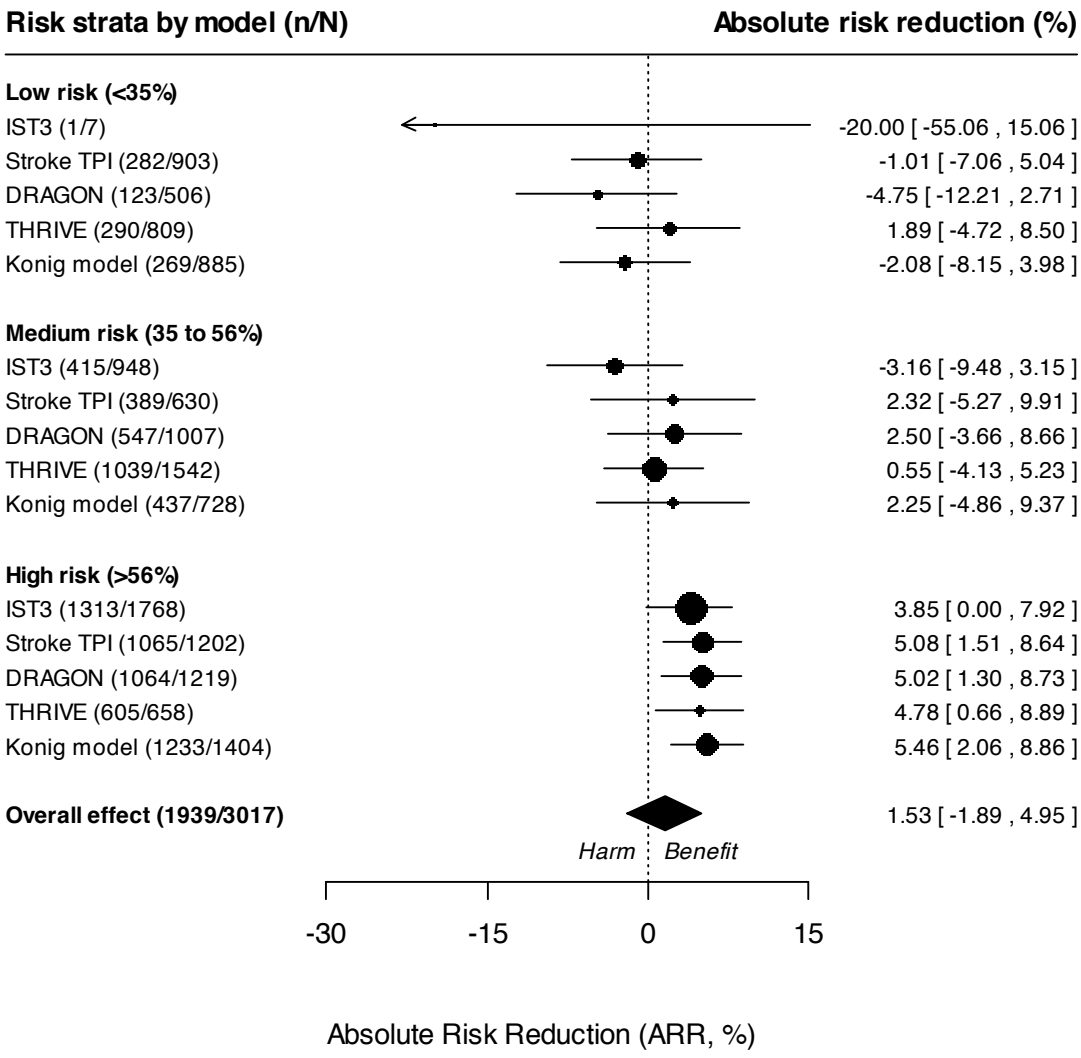


Figure 7-10 Effect of rtPA on six month OHS≥3 stratified by predicted poor functional outcome risk (OHS≥3). Point size is relative to the size of the denominator across risk strata.

7.5.3 Relative risk reduction

A more sensitive test for treatment interaction was explored by modelling the predicted risks directly within a logistic regression (with both binary and ordinal outcomes). The inclusion of an interaction term was assessed via a likelihood ratio test comparing the two nested models: one with the treatment by predicted risk interaction and one without. This formally tests whether or not rtPA acts additively with the log odds of poor functional outcome against predicted risk. An important clinical question is whether or not such an interaction can be classed as *qualitative* or *quantitative*. A qualitative interaction occurs where one group of patients in the trial suffer a qualitatively different treatment effect (e.g., harm) compared to another group (e.g., benefit) with regression slopes going in opposing directions (-ve vs. +ve) across predicted risk.

The dichotomous classification of patients six month functional outcome was defined as an OHS≥3. The ordinal structure of the OHS was also utilised and interpreted as a 5-level outcome collapsing levels 4 to 6 as per the original IST-3 publication (The IST-3 collaborative group, 2012a). This left levels 0, 1, 2, 3 and "4 to 6" as distinct categories whilst still maintaining an ordinal structure. This particular interpretation of the scale has both a clinical and a statistical justification. The clinical justification relates to the qualitative definition of the extent of disability a patient experiences with an OHS score of 4 or 5. Any treatment associated with an improvement at this end of the scale, e.g., moving a patient from an OHS score of 6 to an OHS score of 5, would not be worthwhile. The statistical justification relates to the proportional odds assumption which – under this grouping – becomes more tenable.

Some of the identified prediction models (presented as *risk scores*) only quantified predicted risk at a group level. Those models which grouped the IST-3 patients into three or fewer risk strata levels were included as categorical variables. For those that gave a continuous prediction, or approximately continuous (i.e., >3 levels), the logit transformation was applied and the variable was fit as continuous. The proportional odds assumption was assessed using plots of the score residuals (see section 7.9 Appendix C page 244). The proportional odds assumption seemed plausible with

broadly flat score residuals obtained for each of the included effects. An interaction between a continuous variable, X_1, and a binary variable, X_2, on the log odds scale can be interpreted as a difference in the estimated slope of X_1 given the presence or absence of X_2 (see Chapter 2). The additivity assumption is formally assessed in Table 7-15 with the corresponding interaction plots provided in section 7.9 Appendix C page 244.

Under a proportional odds logistic regression (POLR) model only two out of the 10 presented model fits demonstrated a statistically significant improvement at the 5% level when including the interaction term (i.e., the König model and the IST-3 model). Based on the improvement in the AIC alone interactions between treatment and those models originally derived to predict poor functional outcome post rtPA typically gave a better fit (i.e.: Stroke-TPI [ΔAIC 0.83]; DRAGON [ΔAIC 1.53]; and the König model [ΔAIC 2.22]) in contrast with those models originally derived to predict SICH post rtPA. This may be an artefact of the grouping of predicted risks though which will have caused a loss in precision. Inspection of the interaction plots (see section 7.9 Appendix C on page 244) illustrates that in each case the slopes favour greater benefit (i.e., smaller common ORs<1 for those with a higher predicted risk) which is arguably more quantitative as there is no indication of harm for those at low risk (i.e., no 95%CI with a lower limit greater than one).

Under a binary logistic regression model five out of the 10 model fits suggested a statistically significant interaction at the 5% level. Again this favoured those models originally derived to predict post rtPA poor functional outcome with ΔAIC values ranging from 0.20 to 11.29 (Table 7-15). With reference to the plots provided in Appendix C (section 7.9 on page 244) some of these interactions where qualitative, notably: the Stroke-TPI score (Figure 7-24), the König model (Figure 7-25) and the IST-3 model (Figure 7-26). Caution is stressed in their interpretation though. The POLR was adopted in this analysis as a more sensitive test, however, rather counter intuitively, in some instances the binary logistic regression model reached statistical significance where the POLR did not. To understand why this is it is important to acknowledge the differences in each approach. The POLR relies on the proportional odds assumption which is reasonable with these data meaning that the effect

estimates suggested under the POLR model fits are plausible. In this way the effects are common across all potential dichotomies and in some sense can be viewed as a reasonable average across each split. In the case of the binary logistic regression if other dichotomies are considered (i.e., OHS≥2, OHS≥4 etc.) the qualitative aspect of the interaction effect noted in the binary logistic regression models is attenuated considerably (see section 7.9 Appendix C Figure 7-27, note that only the König model is presented for illustrative purposes).

The results obtained under the POLR models point toward the same position reached in the absolute case. Those at low risk from either post rtPA SICH or poor functional outcome do not appear to derive any overall harm from treatment. Those at a high predicted risk from either event appear to experience more benefit, with a relative treatment effect which may increase with the predicted risk.

Table 7-15 Assessing treatment additivity under binary and ordinal logistic regression fits. Predicted SICH is fitted as a continuous predictor with a logit transform for linearity except where specified (see also section 7.9 Appendix C on page 244).

	Binary logistic regression (OHS≥3)			Proportional odds logistic regression (five levels)		
Model	**ΔAIC[a]**	**χ^2**	**P-value**	**ΔAIC[a]**	**χ^2**	**P-value**
HAT	-0.54	1.46	0.2264	0.20	2.20	0.1384
SEDAN	-0.95	1.05	0.3065	-0.79	1.21	0.2711
SITS	-1.79	0.21	0.6493	-1.98	0.02	0.8837
SPAN-100[b]	3.13	5.13	0.0235	0.28	2.28	0.1312
GRASPS	1.81	3.81	0.0511	-0.59	1.41	0.2359
THRIVE[b,c]	0.20	4.20	0.1222	-1.55	2.45	0.2933
Stroke-TPI	6.70	8.70	0.0032	0.83	2.83	0.0928
DRAGON	6.60	8.60	0.0034	1.53	3.53	0.0602
König model	11.29	13.29	0.0003	2.22	4.22	0.0399
IST3 model	5.09	7.09	0.0078	3.50	5.50	0.0190

a ΔAIC are the difference in AIC units between the interaction model and the additive model. Positive values indicate an improvement in model fit.

b Predicted SICH risk fit as a categorical variable

c LR test compared to a χ^2 with two degrees of freedom.

7.6 Discussion

There was insufficient evidence to support a stratified approach to treating acute ischaemic stroke patients with rtPA according to their predicted risk from harm. There are a number of explanations for this. The discrimination of those with post rtPA SICH events and those without was modest. Theoretically, a better discriminating model could separate those who will suffer a post rtPA SICH from those that will not. As was illustrated in this chapter it would seem that this is not an issue which can be immediately addressed through the use of more sophisticated statistical methodology but will rather involve identifying stronger risk factors through etiological research. It is possible that variables not measured in IST-3, such as advanced imaging methods, genotyping, or blood biomarkers related to the pathophysiology of post-rtPA SICH may better predict response to treatment (Álvarez-Sabín et al., 2013). Based on the current evidence it is unlikely that those available clinical prediction models will play a role in the influence of routine clinical practice.

The correlation between predicted post rtPA SICH and post rtPA poor-functional outcome risk is large with a spearman's correlation coefficient of 0.76 in the IST-3 data. The associations between outcome and predictors are similar regardless of the outcome (Figure 7-11). This explains why the message from Figure 7-9 and from Figure 7-10 was the same.

The position taken in this chapter on the contrasting conclusions drawn from the binary logistic regression and the POLR model (see section 7.5.3 page 229) can be framed as a question of simplicity versus complexity question – first discussed in Chapter 2. Despite the data limitations (note that few patients had a low predicted risk, see section 7.9 Appendix C page 244), the binary logistic model offers the simplest answer to what is arguably a complex question. The POLR adds an extra layer of complexity by requiring the proportional odds assumption to hold, however it offers a halfway-house between the binary fit approach and an approach whereby a different effect is thought to hold for each of the potential dichotomies. The POLR model also makes better use of the available data in a way that the binary logistic

regression does not. The conclusion drawn in this chapter was to favor the POLR model which indicated no excess harm in the low risk patients. This is rather post-hoc though and further investigation within low risk patients is required before evidence of harm could be concluded.

At the time of writing this thesis a collaborative individual patient data meta-analysis between rtPA trialists was underway. The Stroke Thrombolytic Trialists' Collaborative (STTC) used data from all of the available rtPA trials to explore patient subgroups of interest (The Stroke Thrombolysis Trialists' Collaborative Group, 2013). The group published their findings in August 2014. Analysing all of the rtPA trial data, Emberson *et al*. found that the benefit of treatment with rtPA did not depend upon the age or the severity of the patient, concluding benefit for older patients as well as those with milder strokes (Emberson et al., 2014). They also emphasised the urgency in onset time to treatment illustrating that the effect of rtPA in improving the odds of a good outcome reduced the later it was given.

The direction of the results and the conclusions drawn were not affected by adopting various sensitivity analyses (see section 7.7 Appendix A on page 236)

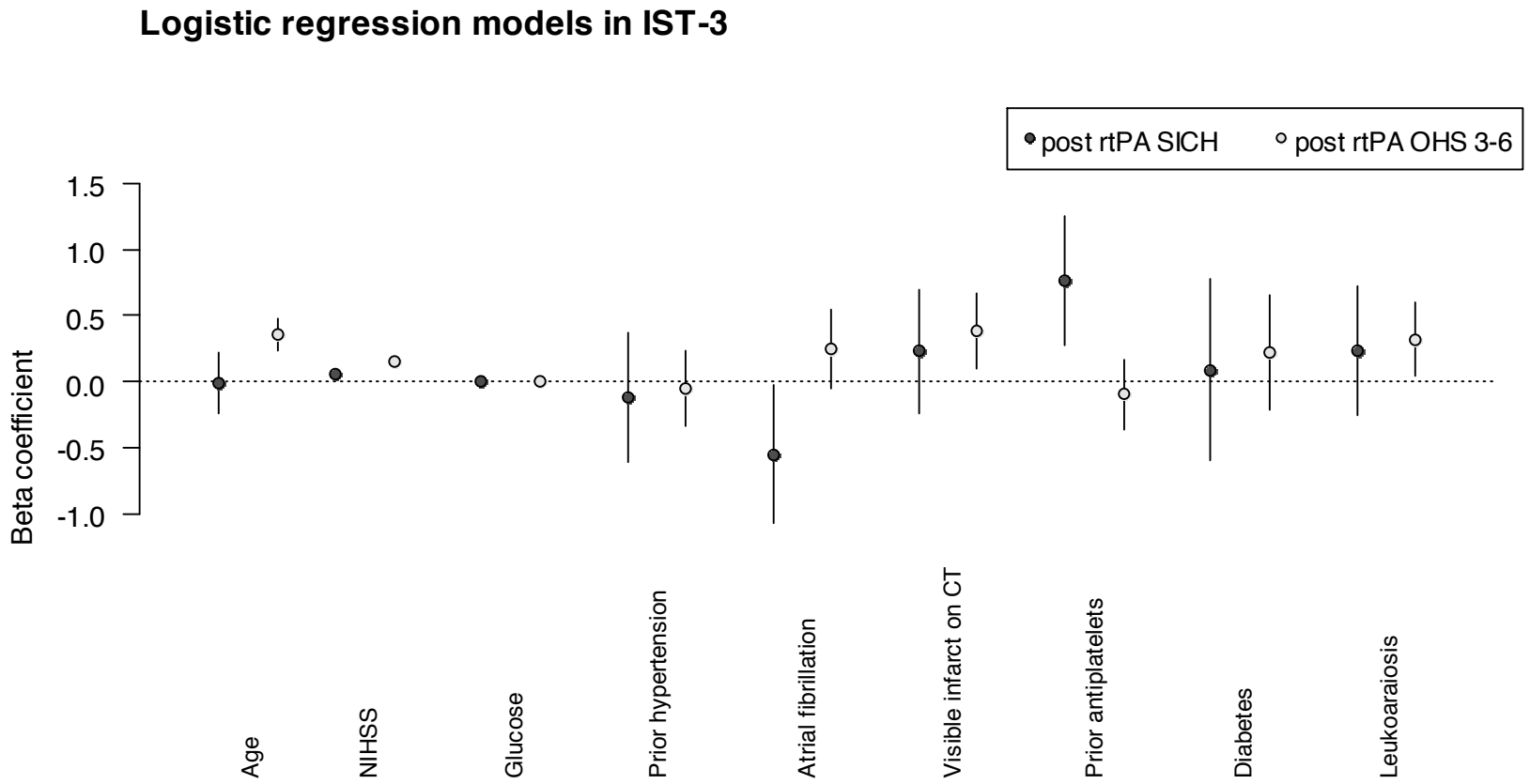


Figure 7-11 Comparing effect sizes between post rtPA SICH and OHS≥3 as estimated using the IST-3 dataset.

7.7 Appendix A: Sensitivity analyses and additional tables

The various sensitivity analyses conducted in this chapter are summarised in Table 7-16 below. Additional tables summarising formal comparisons of areas under the ROC curve are given for post rtPA: (i) SICH (Table 7-17); parenchymal haemorrhage (Table 7-18); and OHS≥3 (Table 7-19).

Table 7-16 Summary of sensitivity analyses.

	Per protocol rtPA randomisation	**Randomised to treatment within 4.5 hours**	**Randomised to treatment after 4.5 hours**
Analysis	**Comment**	**Comment**	**Comment**
Model development			
	A reduction of 25 patients. No difference in direction of covariate effects or effect size, slight inflation in SE as expected	A reduction of 460 patients. No difference in direction of covariates, but slight differences in effect size. Large inflation of SE. Likely due to reduction in sample size	A reduction of 901 patients. Diabetes becomes protective. Hypertension significantly increases patient risk of SICH. It is noted that when simply adjusting for time to treatment, there is no significant effect
Model evaluation			
Parenchymal haemorrhage	No large quantitative differences in model calibration or discrimination. No differences in nonparametric comparisons of ROC curves	Reduction in sample size is expressed in the pseudo R^2 values. Differences in nonparametric comparison of ROC curves: likely due to reduction in sample size	No difference, small differences resulting from reduced sample; differences in discrimination, balanced by larger 95% CI intervals
SICH	As above	As above	As above. Some large differences in R^2 noted
OHS≥3	As above	As above. Slight improvement in discrimination, but balanced by an increase in the width of the 95% CIs. Differences in nonparametric comparison of ROC curves: likely due to reduction in sample size	As above. Differences in discrimination noted, likely accountable to sample size.

Table 7-17 Assessing the prognostic value of different clinical prediction models for post rtPA SICH and poor functional outcome by nonparametric comparison of ROC curves computed using the IST3 data for post rtPA SICH.

Model	HAT	SEDAN	SITS	GRASPS	SPAN-100	STROKE TPI	DRAGON	THRIVE	König model
HAT	-	-	-	-	-	-	-	-	-
SEDAN	0.2,.8831	-	-	-	-	-	-	-	-
SITS	0.6, .5513	0.5, .6353	-	-	-	-	-	-	-
GRASPS	0.2, .8558	0.0, .9669	-0.5, .6426	-	-	-	-	-	-
SPAN-100	-1.6, .1051	-1.6, .1126	-2.1, .0392	-2.3, .0222	-	-	-	-	-
STROKE TPI	0.8, .4211	0.5, .6310	-0.2, .8347	0.5, .5940	2.8, .0051	-	-	-	-
DRAGON	1.3, .2099	1.2, .2410	0.2, .8617	1.0, .3167	2.6, .0097	0.7, .5066	-	-	-
THRIVE	-0.5, .6557	-0.6, .5694	-1.0, .3034	-1.1, .2592	1.9, .0636	-1.3, .1995	-1.5, .1234	-	-
König model	0.3, .7665	0.1, .9077	-0.4, .6976	0.1, .8873	3.4, .0007	-0.6, .5396	-0.9, .3494	1.4, .8576	-
IST3 model	2.8, .0047	2.5, .0123	1.9, .0557	2.4, .0179	3.9, .0001	2.4, .0159	1.9, .0592	2.8, .0045	2.6, .0106

NOTE: Data are Z-statistics and P values relating to the pairwise nonparametric comparison of AUROCC values

Table 7-18 Assessing the prognostic value of different clinical prediction models for post rtPA SICH and poor functional outcome by nonparametric comparison of ROC curves computed using the IST3 data for post rtPA parenchymal haemorrhage.

Model	HAT	SEDAN	SITS	GRASPS	SPAN-100	STROKE TPI	DRAGON	THRIVE	König model
HAT	-	-	-	-	-	-	-	-	-
SEDAN	0.0, .8506	-	-	-	-	-	-	-	-
SITS	0.7, .4130	0.7, .4035	-	-	-	-	-	-	-
GRASPS	0.1, .7091	0.0, .8634	0.6, .4500	-	-	-	-	-	-
SPAN-100	7.4, .0066	5.7, .0174	2.6, .1071	9.1, .0026	-	-	-	-	-
STROKE TPI	0.1, .7915	0.2, .6879	1.1, .2931	0.7, .4068	15.4, .0001	-	-	-	-
DRAGON	1.3, .2575	2.6, .1104	2.9, .0897	2.8, .0961	15.3, .0001	1.5, .2260	-	-	-
THRIVE	0.2, .6819	0.1, .7831	0.4, .5297	0.1, .8150	10.5, .0012	0.7, .4178	2.4, .1236	-	-
König model	0.0, .9050	0.1, .8043	1.0, .3188	0.6, .4384	22.8, <.0001	0.0, .8467	1.2, .2665	0.1, .8018	-
IST3 model	0.7, .4038	0.8, .3824	3.6, .0590	1.2, .2814	11.5, <.0001	0.5, .4666	0.0, .9910	1.2, .2752	0.6, .4290

NOTE: Data are Z-statistics and P values relating to the pairwise nonparametric comparison of AUROCC values

Table 7-19 Assessing the prognostic value of different clinical prediction models for post rtPA SICH and poor functional outcome by nonparametric comparison of ROC curves computed using the IST3 data for post rtPA death or dependency.

Model	HAT	SEDAN	SITS	GRASPS	SPAN-100	STROKE TPI	DRAGON	THRIVE	König model
HAT	-	-	-	-	-	-	-	-	-
SEDAN	2.8, .0048	-	-	-	-	-	-	-	-
SITS	-2.7, .0062	-5.0, <.0001	-	-	-	-	-	-	-
GRASPS	4.4, <.0001	2.3, .0233	7.5, <.0001	-	-	-	-	-	-
SPAN-100	-3.3, .0010	-5.3, <.0001	0.4, .7286	-8.6, <.0001	-	-	-	-	-
STROKE TPI	8.7, <.0001	6.7, <.0001	10.0, <.0001	5.0, <.0001	13.5, <.0001	-	-	-	-
DRAGON	6.9, <.0001	6.0, <.0001	8.3, <.0001	3.4, .0663	9.9, <.0001	-2.9, .0038	-	-	-
THRIVE	4.4, <.0001	2.3, .0216	7.6, <.0001	0.1, .9000	9.6, <.0001	-4.6, <.0001	-1.6, .1035	-	-
König model	7.6, <.0001	5.4, <.0001	9.5, <.0001	4.8, <.0001	14.3, <.0001	-0.7, .5086	2.2, .0312	4.8, .0168	-
IST3 model	0.3, .7965	-2.4, .0165	3.7, .0003	-4.4, <.0001	3.1, .00182	-8.1, <.0001	-6.4, <.0001	-4.2, <.0001	-7.5, <.0001

NOTE: Data are Z-statistics and P values relating to the pairwise nonparametric comparison of AUROCC values

7.8 Appendix B: Re-calibration in IST-3 dataset

The calibration metrics for the Stroke-TPI score, the Konig model and the IST-3 model are provided in Table 7-20. In each case the estimated correction factor (*cf*) represents the log odds ratio of incidence and estimated risk. Recalibration involves adding the *cf* to in each instance to the linear predictor therefore adjusting for the observed differences in incidence of each outcome in IST-3. The benefit of recalibration is illustrated by comparing the updated calibration metrics in Table 7-20 with the original metrics presented in Table 7-5. For instance, recalibration considerably improved the systematic under-prediction of patients risk from post rtPA poor functional outcome according to the Stroke-TPI score, with a calibration intercept that was 2.49 reduced to 0.68. Note that Figure 7-12, Figure 7-13 and Figure 7-14 illustrate the impact of recalibration in each case.

Table 7-20 Models recalibrated for IST-3 data for: (i) post rtPA SICH; and (ii) post rtPA poor functional outcome (OHS≥3). Note intervals provided are 95%CIs.

Model	SICH	OHS≥3
Stroke-TPI		
Correction factor (*cf*)	-1.39	1.81
events/Total	87/1365	856/1365
Calibration intercept (95%CI)	-0.34 (-0.57 to -0.10)	0.68 (0.55 to 0.81)
Calibration slope (95%CI)	0.33 (0.17 to 0.48)	0.99 (0.87 to 1.11)
König model		
Correction factor (*cf*)	-3.32	-0.14
events/Total	102/1507	957/1507
Calibration intercept (95%CI)	-1.14 (-1.36 to -0.92)	0.17 (0.05 to 0.30)
Calibration slope (95%CI)	0.27 (0.14 to 0.40)	0.84 (0.74 to 0.94)
IST-3 model		
Correction factor (*cf*)	-	3.26
events/Total	-	957/1507
Calibration intercept (95%CI)	-	0.11 (0.00 to 0.22)
Calibration slope (95%CI)	-	1.71 (1.44 to 1.98)

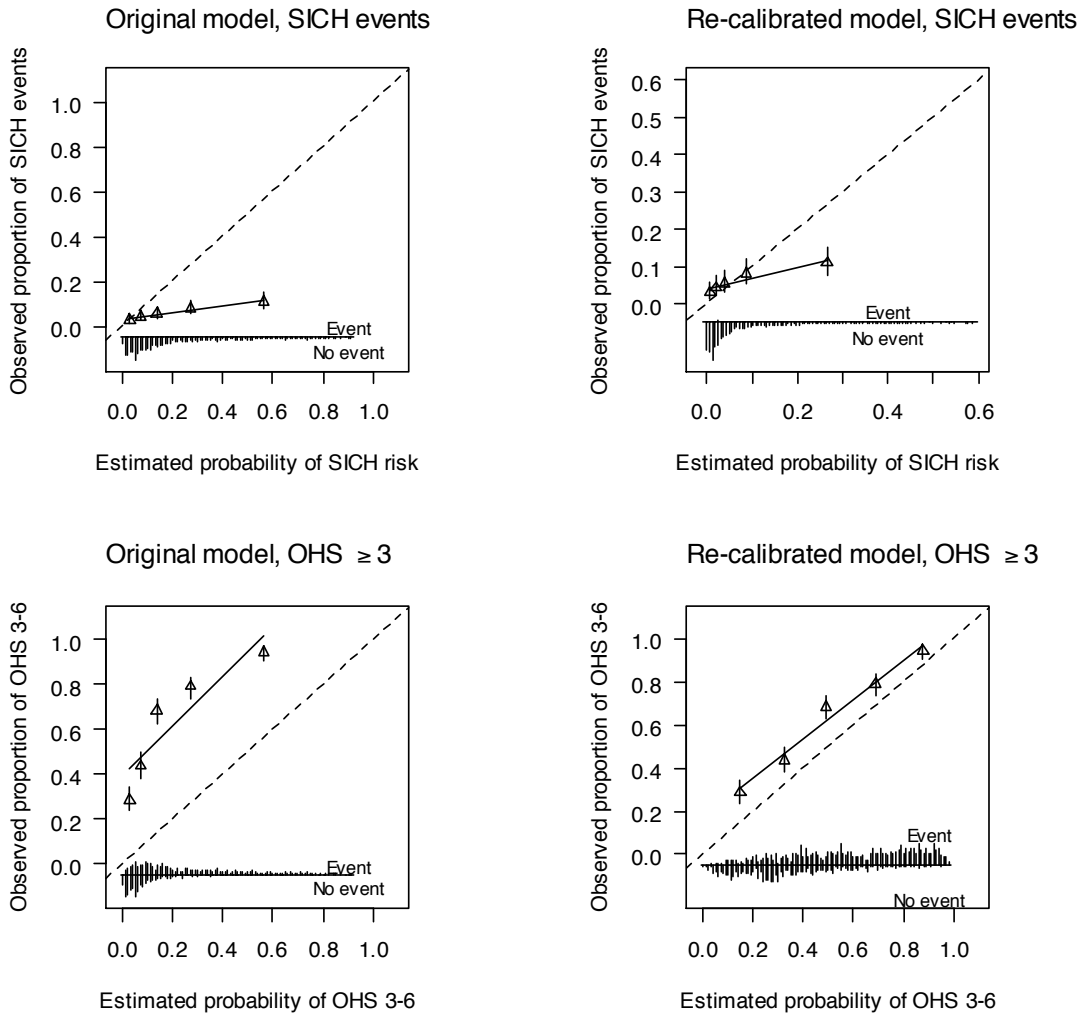


Figure 7-12 Original and updated calibration plots for Stroke-TPI model in IST-3 dataset.

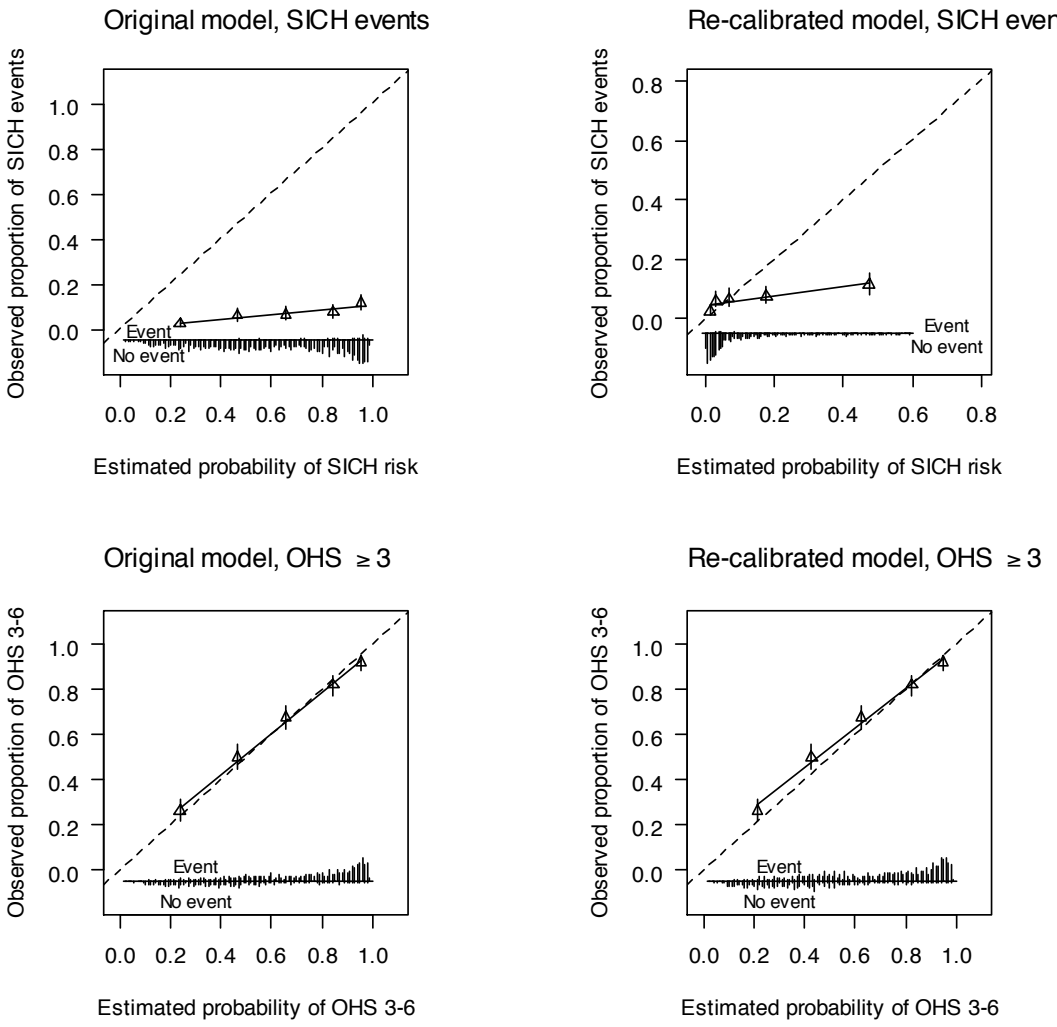


Figure 7-13 Original and updated calibration plots for König model in IST-3 dataset.

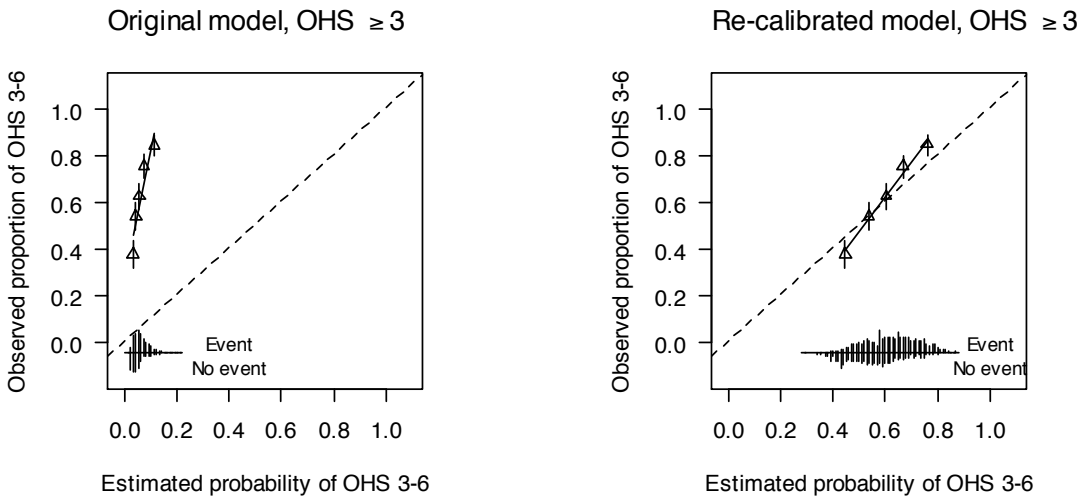


Figure 7-14 Original and updated calibration plots for IST-3 model in IST-3 dataset.

7.9 Appendix C: Interaction on relative risk scale

A formal test for a *treatment by predicted risk* interaction is presented here on the relative risk scale. This was undertaken using the OHS outcome as a binary outcome (OHS≥3) as well as when maintaining the full ordinal structure of the OHS outcome (as defined previously).

The proportional odds assumption is assessed visually using the residual plots provided below (Figure 7-16, Figure 7-17 and Figure 7-18). The largely flat residuals support the proportional odds assumption in each case.

Where possible prediction models were recalibrated (i.e., Stroke-TPI (Figure 7-12) and the model proposed by König *et al*. (Figure 7-13) see Appendix B above). This improves the obtained predicted risks which, after recalibration, adequately reflect the observed difference in the incidence of post rtPA SICH as recorded in IST-3 in contrast to the incidence noted in development.

In each of the interaction plots provided below (see Figure 7-19 to Figure 7-26) the left hand-side shows the interaction from a binary logistic regression model whilst the right hand-side shows the interaction from a POLR model. See results section 7.5.3 on page 229 for a full discussion.

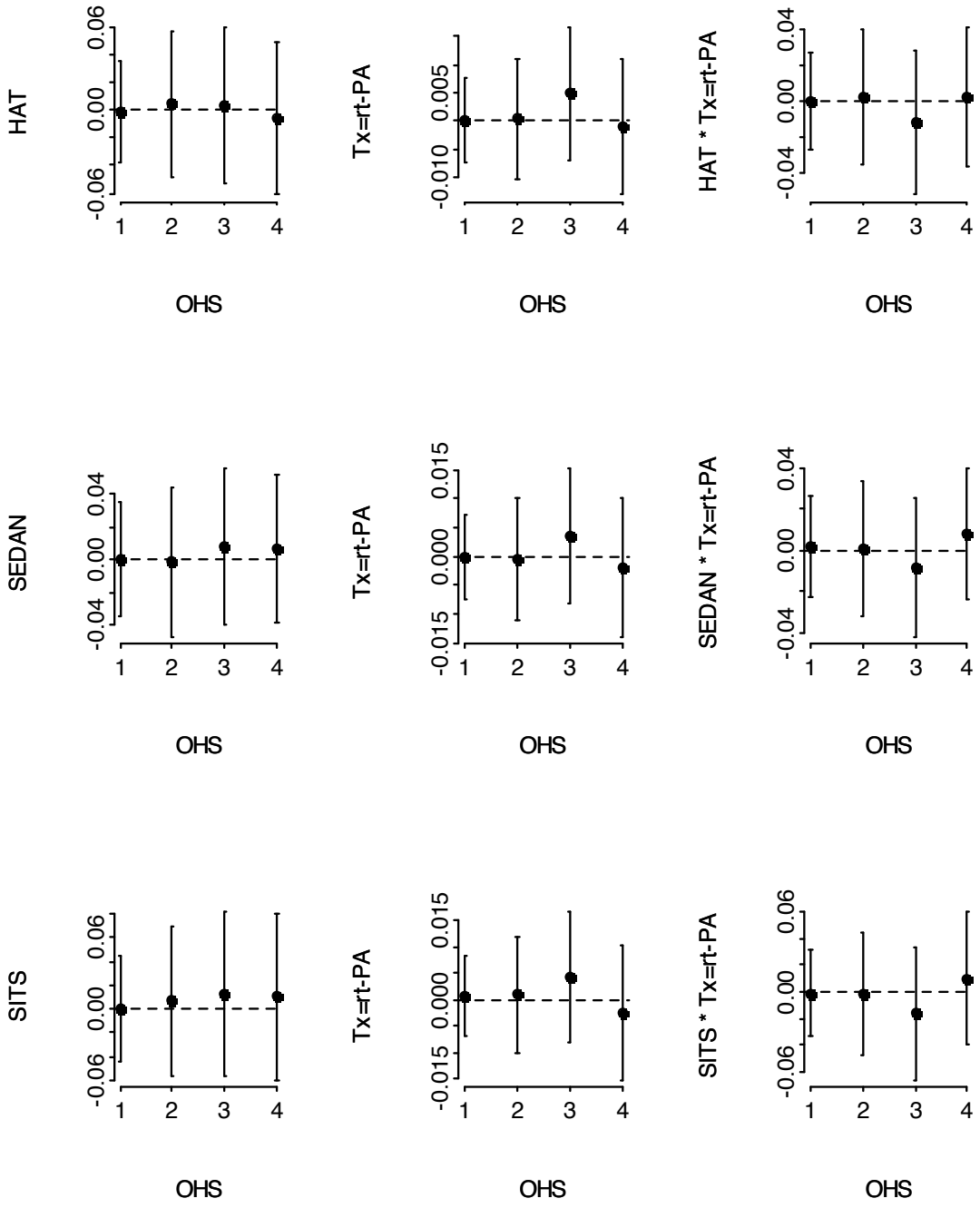


Figure 7-15 Score residual plots for HAT, SEDAN and SITS (binary model score residuals for all cut-points of the ordinal outcome). Each model includes the predicted risk, the treatment variable and the interaction of the two.

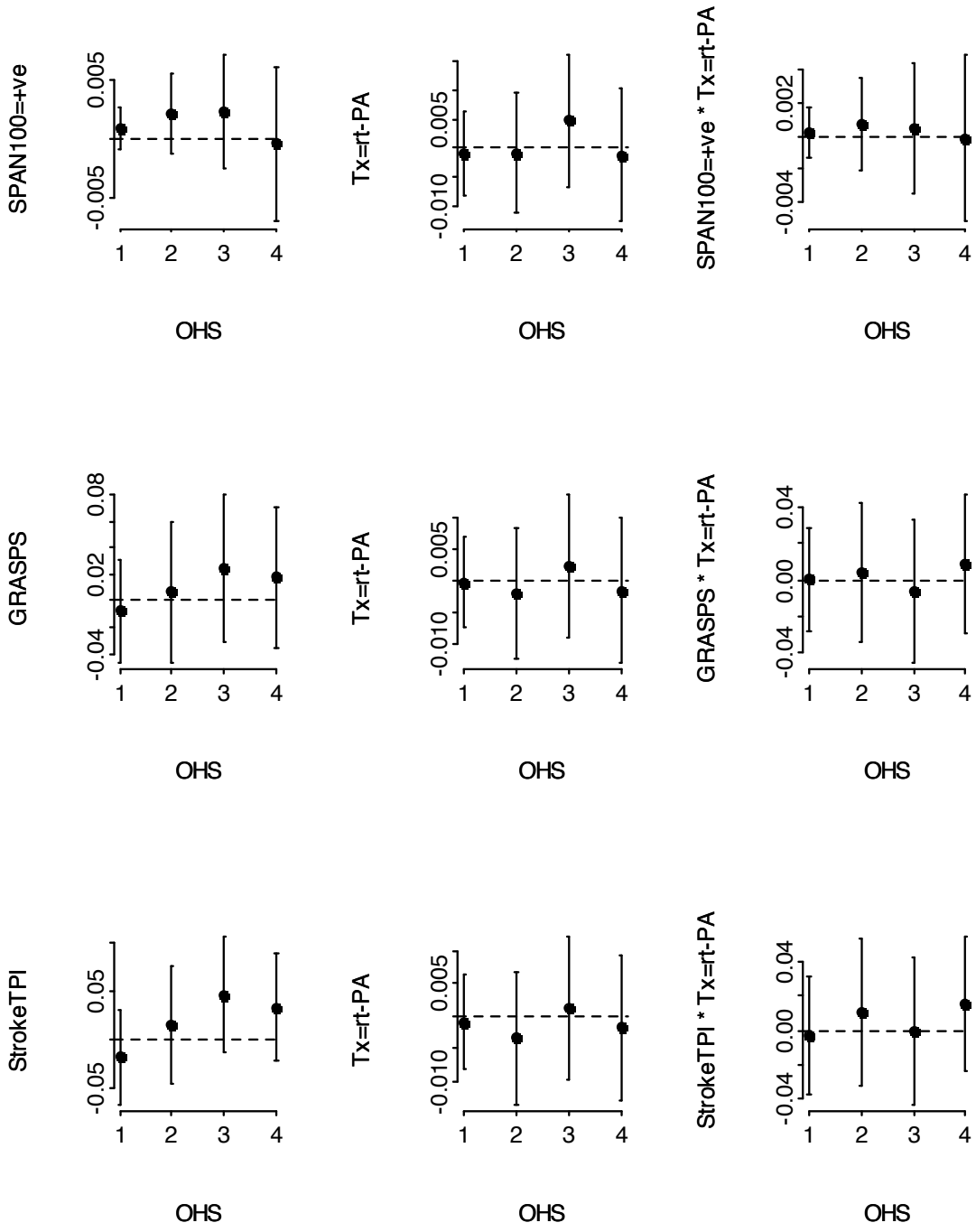


Figure 7-16 Score residual plots for SPAN-100, GRASPS and Stroke-TPI (binary model score residuals for all cut-points of the ordinal outcome). Each model includes the predicted risk, the treatment variable and the interaction of the two.

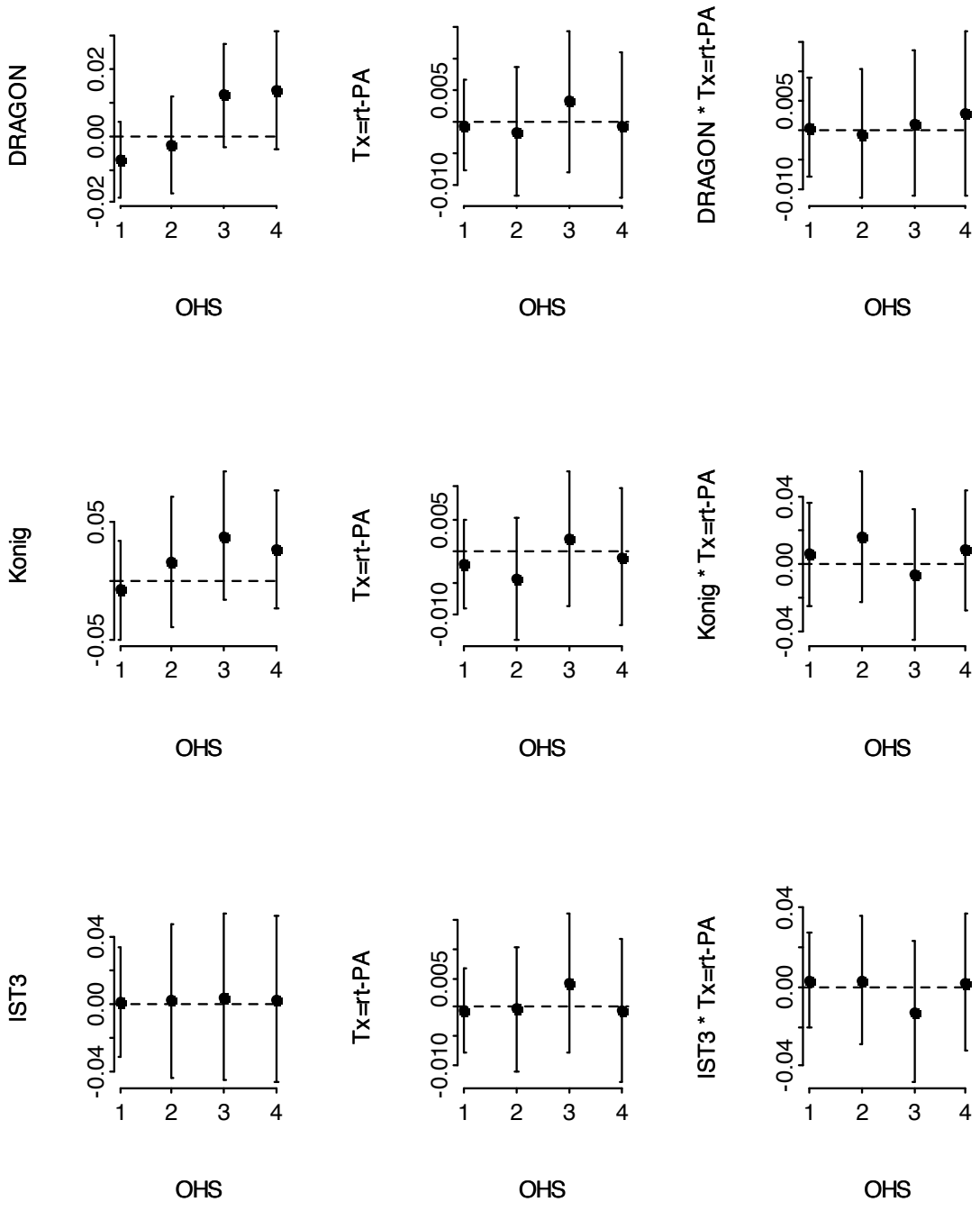


Figure 7-17 Score residual plots for DRAGON, König model and the IST-3 model and SITS (binary model score residuals for all cut-points of the ordinal outcome). Each model includes the predicted risk, the treatment variable and the interaction of the two.

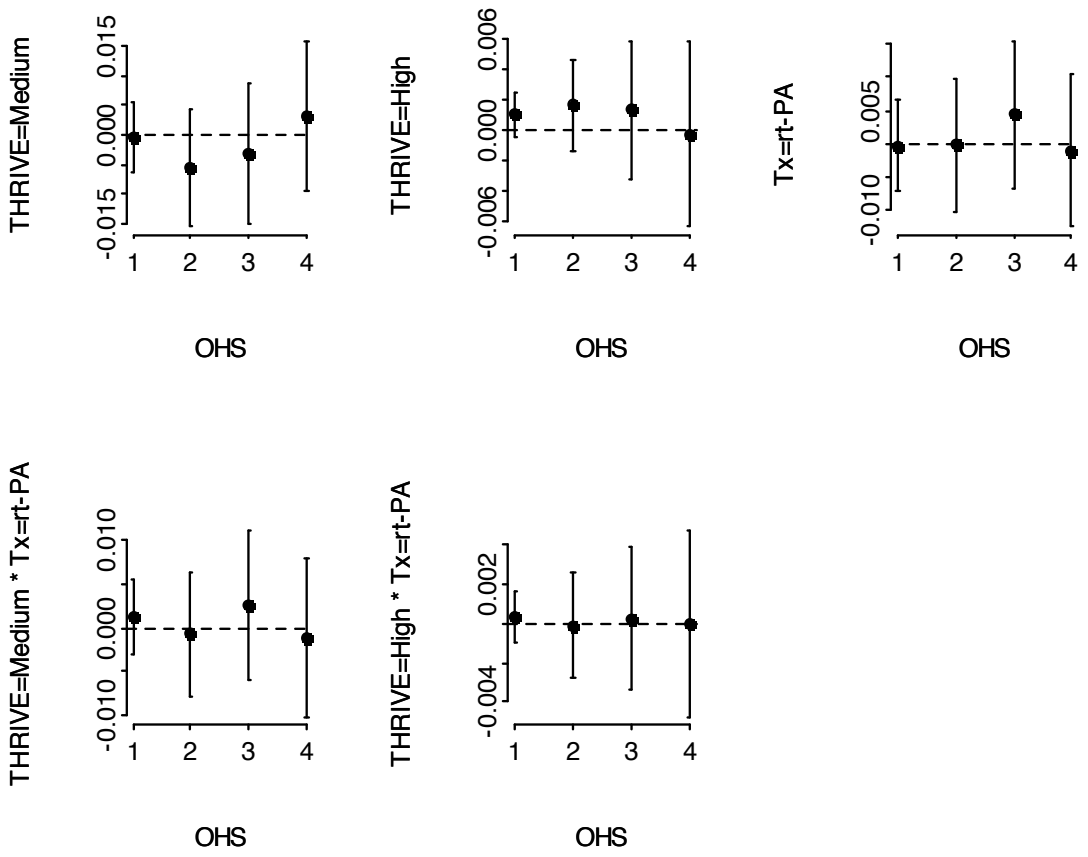


Figure 7-18 Score residual plots for the THRIVE score (binary model score residuals for all cut-points of the ordinal outcome). Predicted risk is included as categorical variable, the treatment variable and the interaction of the two.

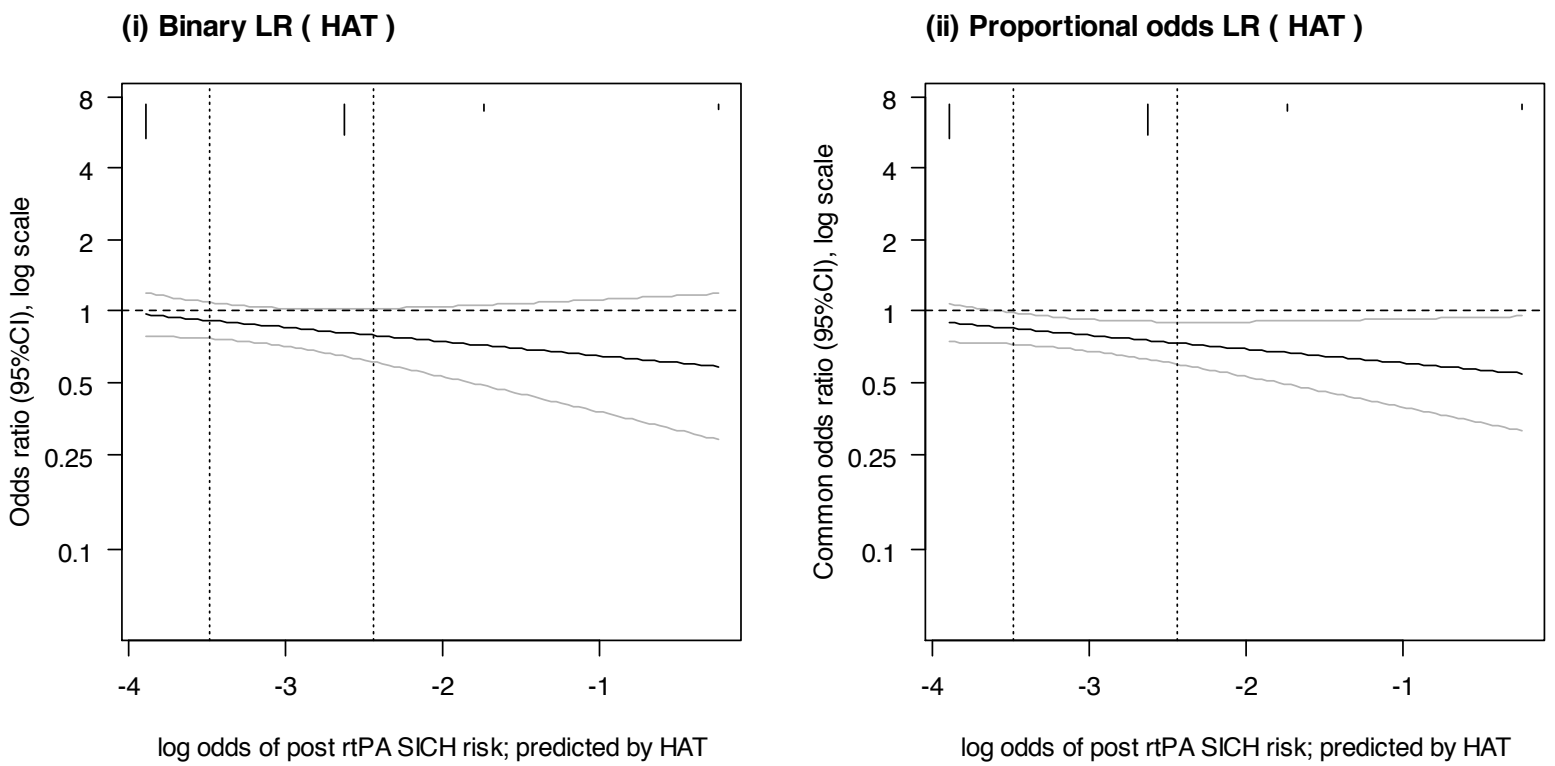


Figure 7-19 The effect of rtPA on six month functional outcome and patients predicted risk of post rtPA SICH by the HAT model: (i) a poor functional outcome; and (ii) disability or death across a five levelled categorisation of OHS stroke outcome. A global test of additivity (rtPA with predicted SICH risk) gave P-values of 0.2264 and 0.1384 for a binary LR (i) and a POLR (ii) respectively. Vertical dashed lines correspond to the categorisation of risk used in the assessment of the absolute risk reduction of rtPA.

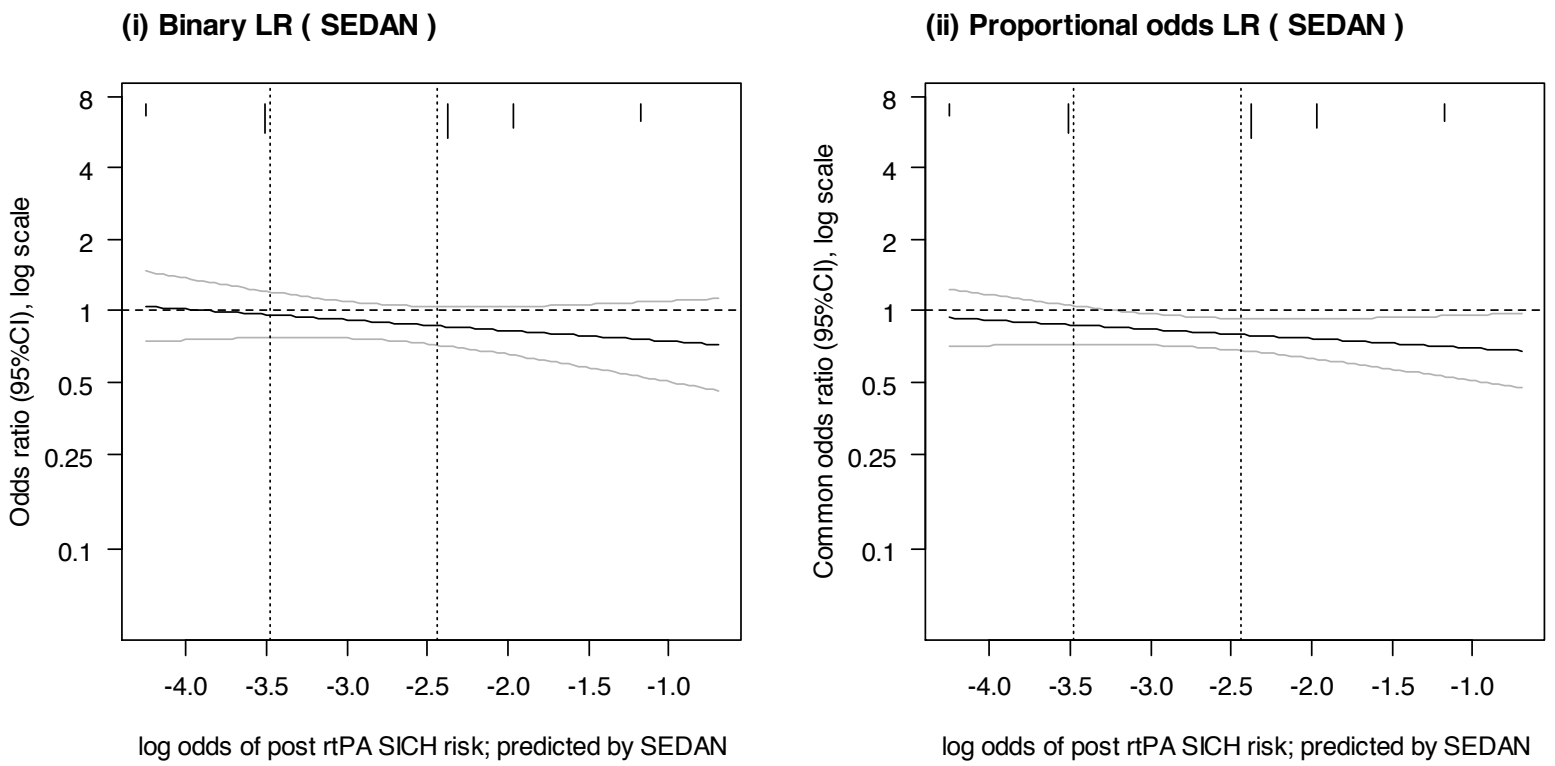


Figure 7-20 The effect of rtPA on six month functional outcome and patients predicted risk of post rtPA SICH by the SEDAN model: (i) a poor functional outcome; and (ii) disability or death across a five levelled categorisation of OHS stroke outcome. A global test of additivity (rtPA with predicted SICH risk) gave P-values of 0.3065 and 0.2711 for a binary LR (i) and a POLR (ii) respectively. Vertical dashed lines correspond to the categorisation of risk used in the assessment of the absolute risk reduction of rtPA.

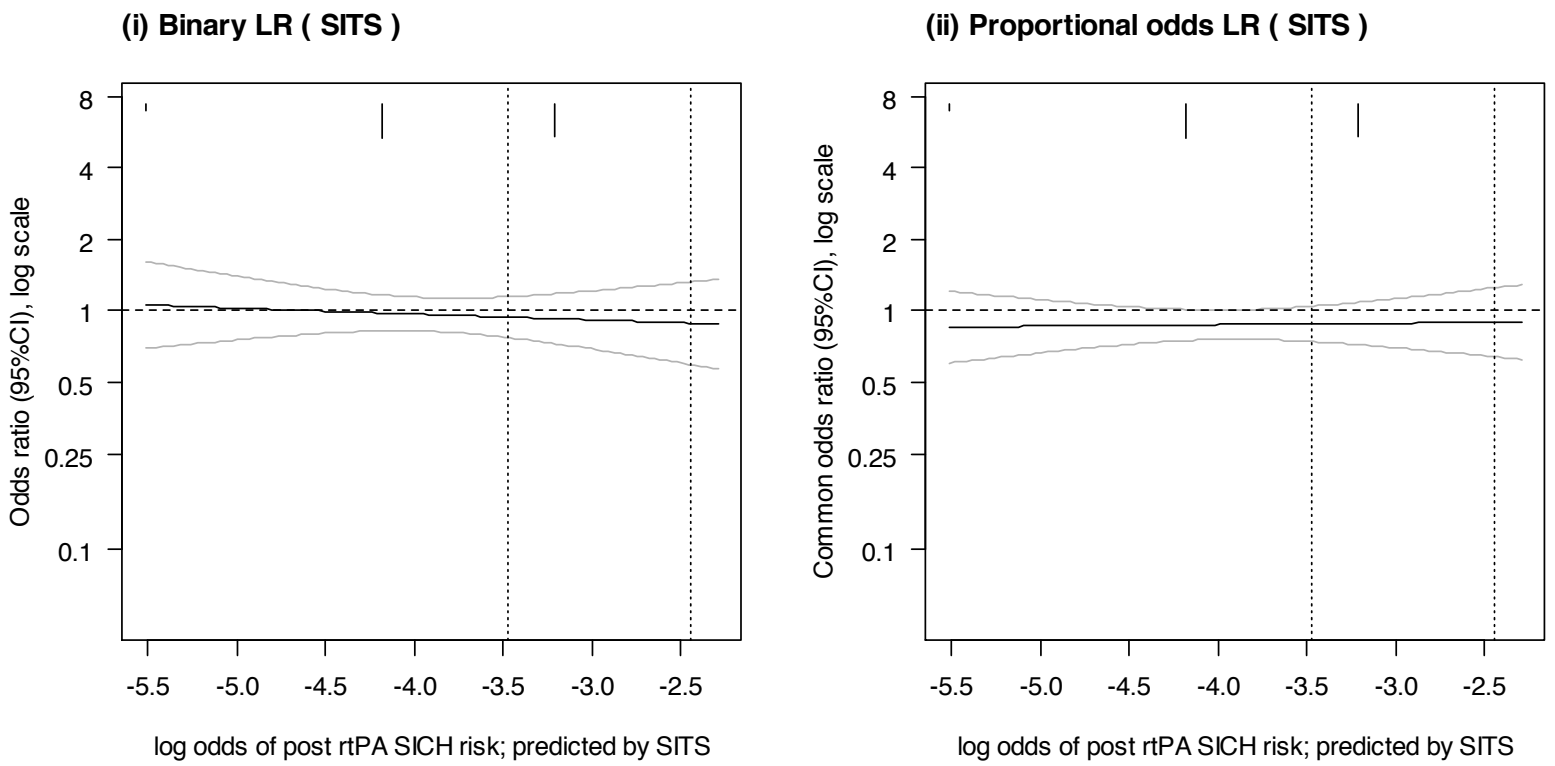


Figure 7-21 The effect of rtPA on six month functional outcome and patients predicted risk of post rtPA SICH by the SITS model: (i) a poor functional outcome; and (ii) disability or death across a five levelled categorisation of OHS stroke outcome. A global test of additivity (rtPA with predicted SICH risk) gave P-values of 0.6493 and 0.8837 for a binary LR (i) and a POLR (ii) respectively. Vertical dashed lines correspond to the categorisation of risk used in the assessment of the absolute risk reduction of rtPA.

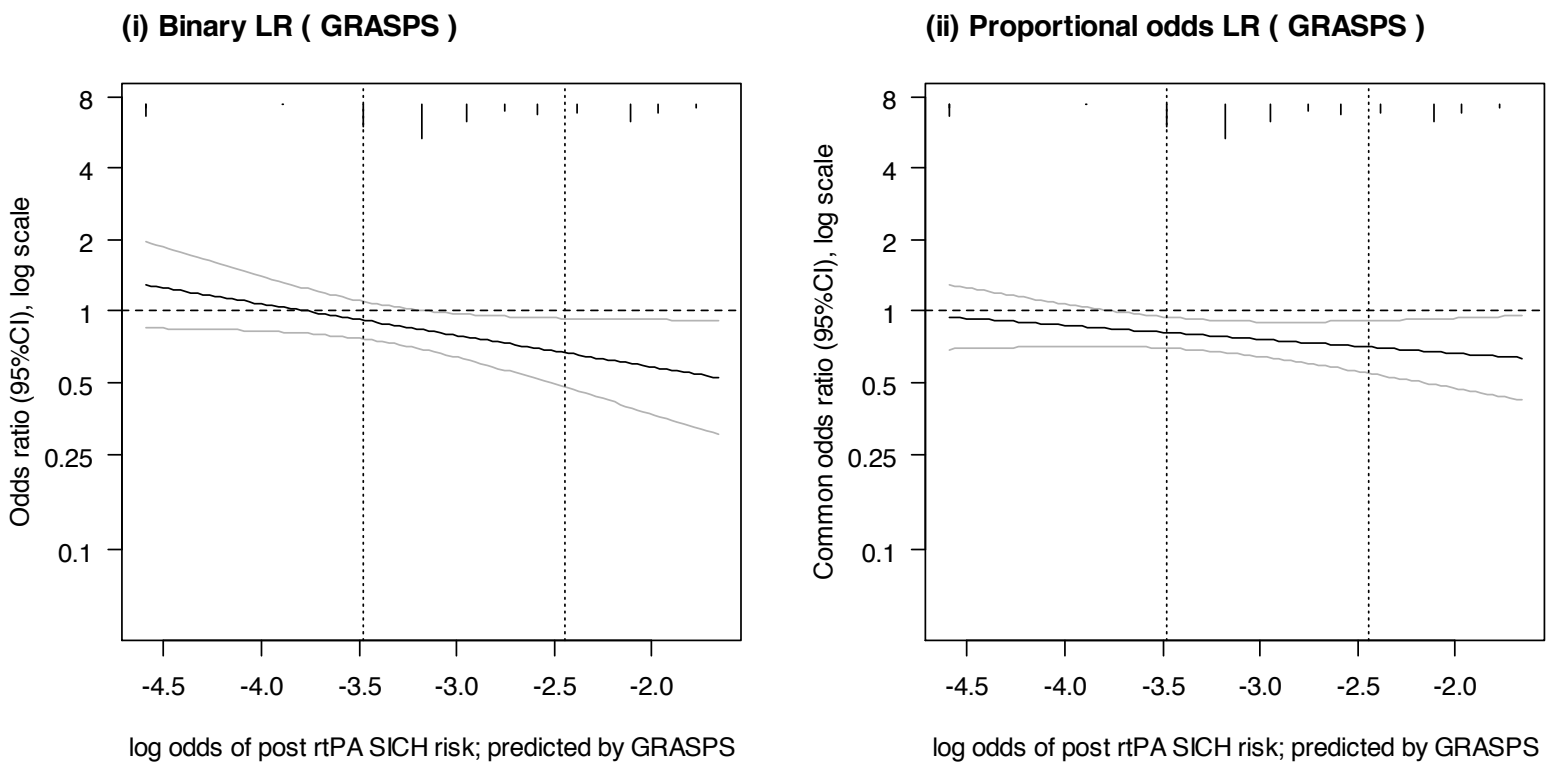


Figure 7-22 The effect of rtPA on six month functional outcome and patients predicted risk of post rtPA SICH by the GRASPS model: (i) a poor functional outcome; and (ii) disability or death across a five levelled categorisation of OHS stroke outcome. A global test of additivity (rtPA with predicted SICH risk) gave P-values of 0.0511 and 0.2359 for a binary LR (i) and a POLR (ii) respectively. Vertical dashed lines correspond to the categorisation of risk used in the assessment of the absolute risk reduction of rtPA.

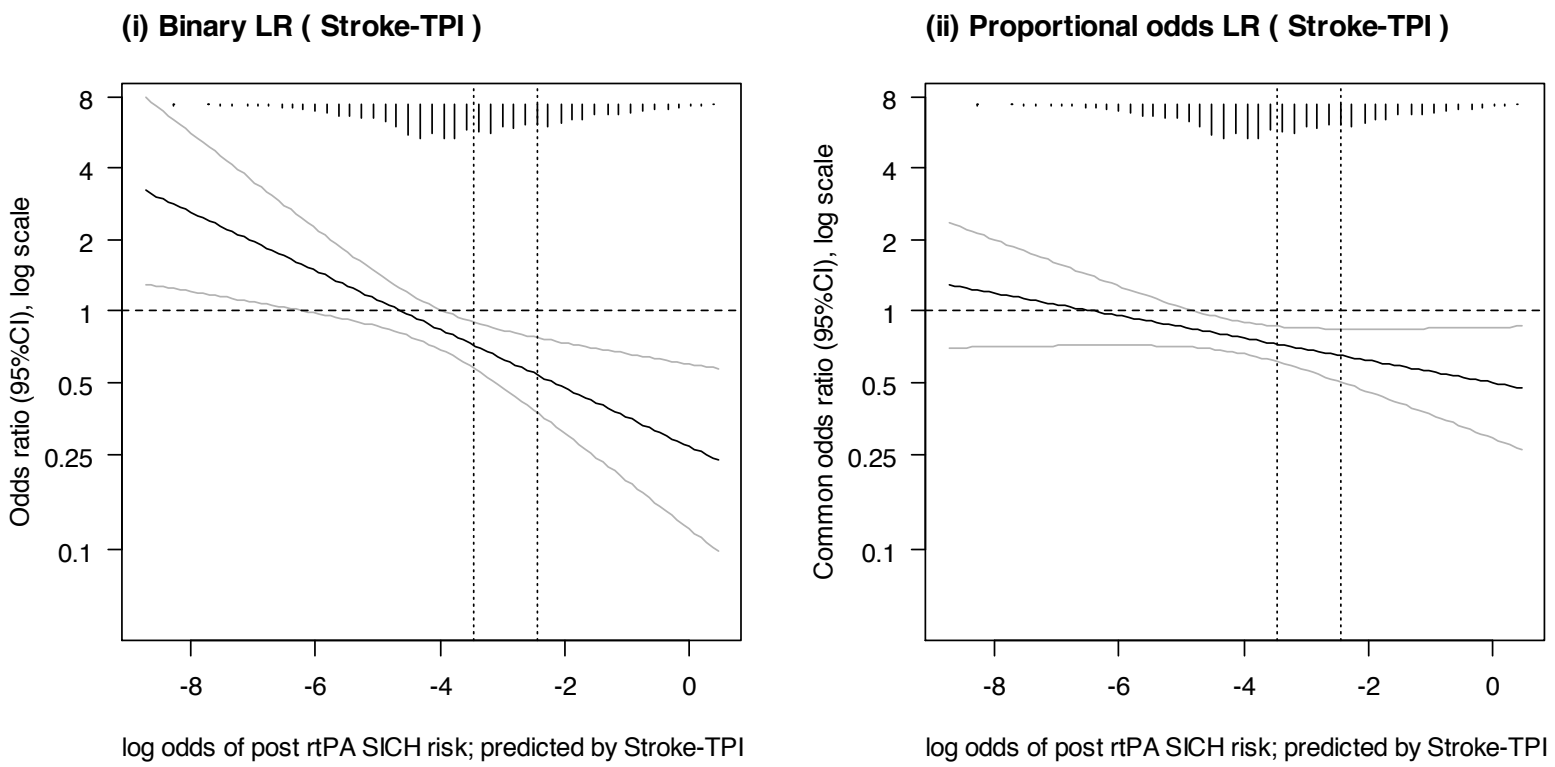


Figure 7-23 The effect of rtPA on six month functional outcome and patients predicted risk of post rtPA SICH by the Stroke-TPI model: (i) a poor functional outcome; and (ii) disability or death across a five levelled categorisation of OHS stroke outcome. A global test of additivity (rtPA with predicted SICH risk) gave P-values of 0.0032 and 0.0928 for a binary LR (i) and a POLR (ii) respectively. Vertical dashed lines correspond to the categorisation of risk used in the assessment of the absolute risk reduction of rtPA.

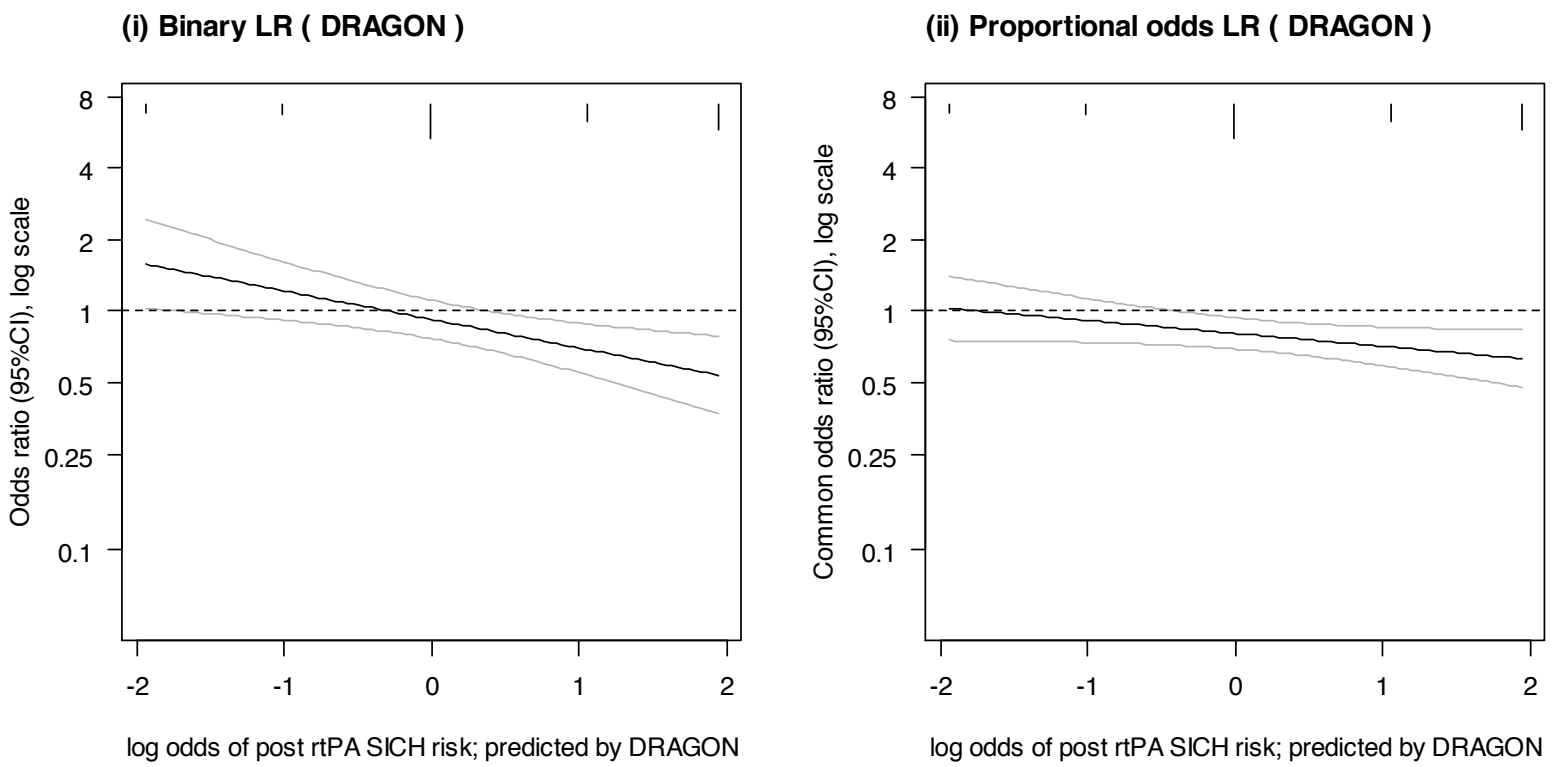


Figure 7-24 The effect of rtPA on six month functional outcome and patients predicted risk of post rtPA SICH by the DRAGON score: (i) a poor functional outcome; and (ii) disability or death across a five levelled categorisation of OHS stroke outcome. A global test of additivity (rtPA with predicted SICH risk) gave P-values of 0.0034 and 0.0602 for a binary LR (i) and a POLR (ii) respectively

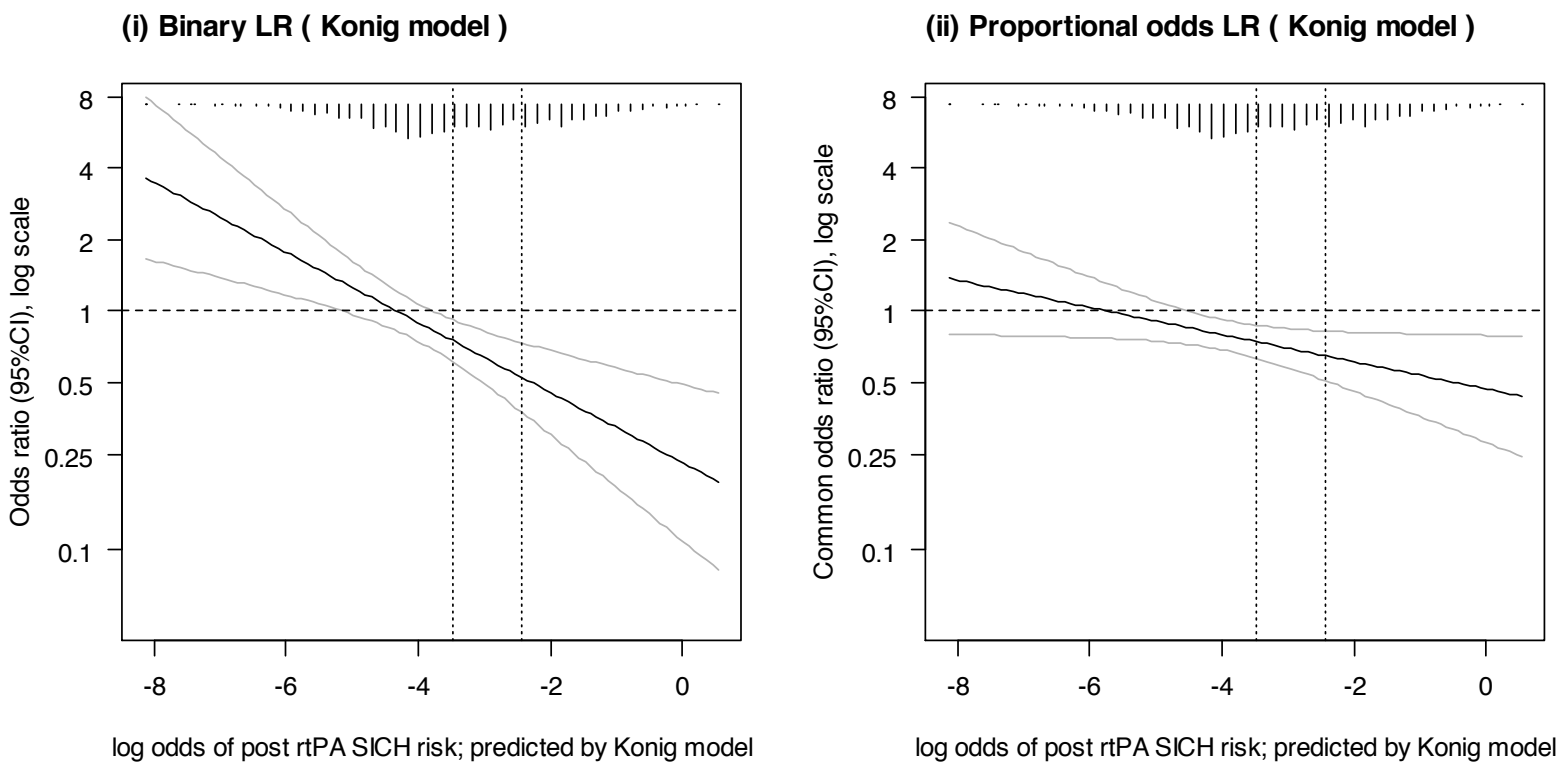


Figure 7-25 The effect of rtPA on six month functional outcome and patients predicted risk of post rtPA SICH by the König model: (i) a poor functional outcome; and (ii) disability or death across a five levelled categorisation of OHS stroke outcome. A global test of additivity (rtPA with predicted SICH risk) gave P-values of 0.0003 and 0.0399 for a binary LR (i) and a POLR (ii) respectively. Vertical dashed lines correspond to the categorisation of risk used in the assessment of the absolute risk reduction of rtPA.

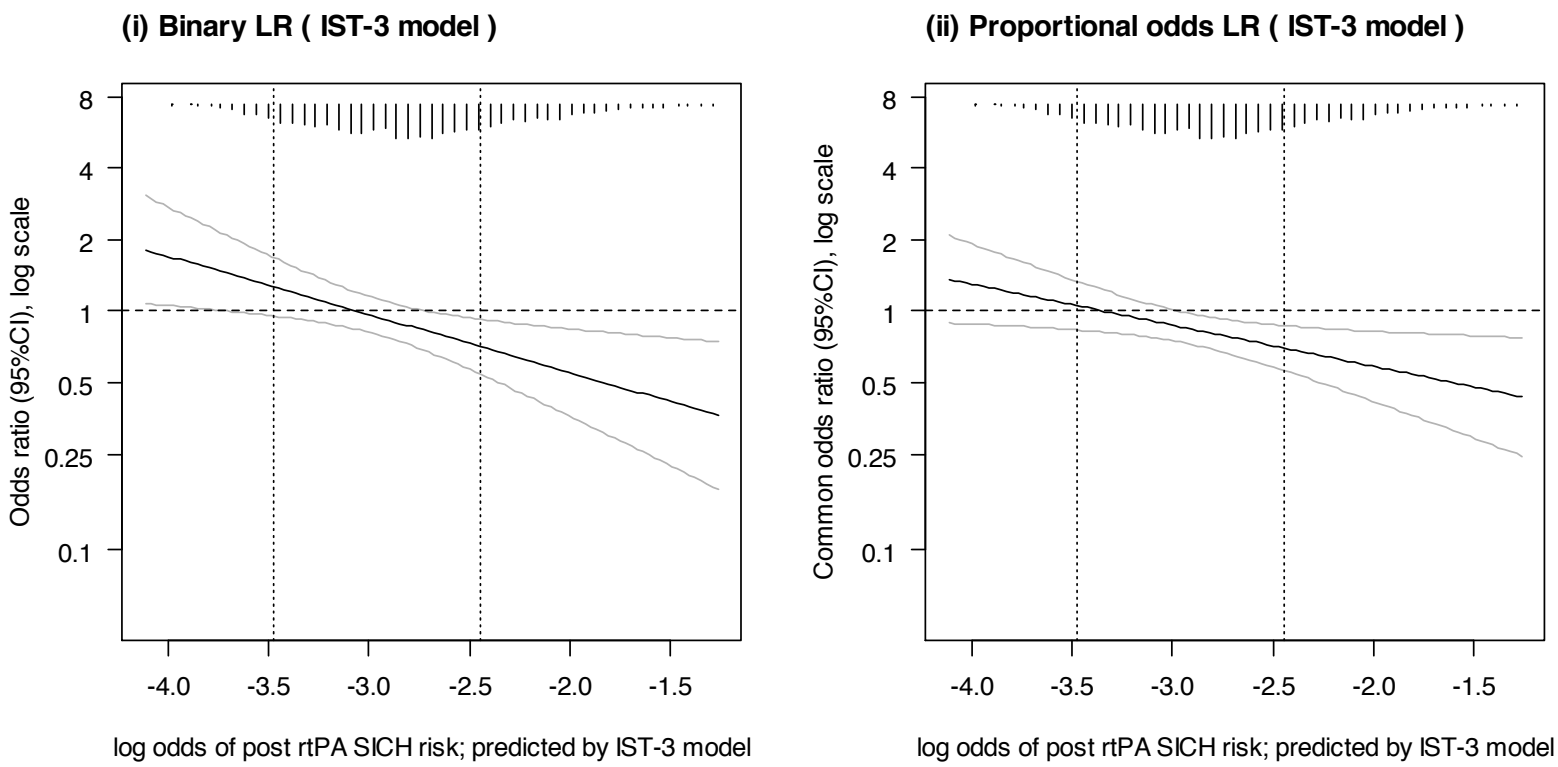


Figure 7-26 The effect of rtPA on six month functional outcome and patients predicted risk of post rtPA SICH by the internally developed IST-3 model: (i) a poor functional outcome; and (ii) disability or death across a five levelled categorisation of OHS stroke outcome. A global test of additivity (rtPA with predicted SICH risk) gave P-values of 0.0078 and 0.0190 for a binary LR (i) and a POLR (ii) respectively. Vertical dashed lines correspond to the categorisation of risk used in the assessment of the absolute risk reduction of rtPA.

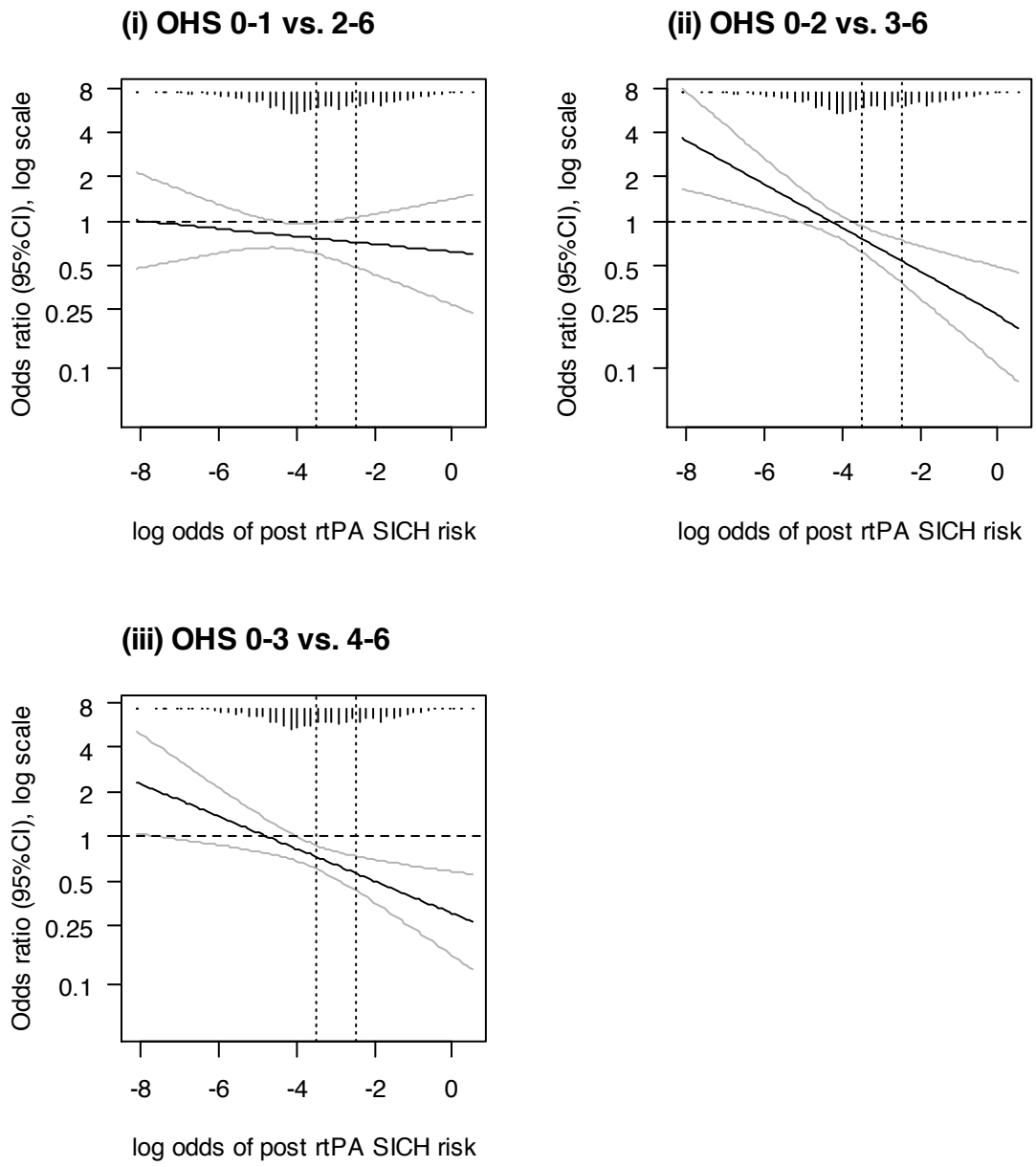


Figure 7-27 The effect of rtPA on six month functional outcome and patients predicted risk of post rtPA SICH by the König model: (i) a dichotomy set at OHS≥2; (ii) a dichotomy set at OHS≥3; and (iii) a dichotomy set at OHS≥4.

Chapter 8: The effect of rtPA on mortality

Background and summary

Treatment with rtPA as early as possible after an acute ischaemic stroke can reduce the severity of six month disability. The impact on long term survival is less clear. In this chapter a secondary analysis of the 18 month IST-3 data is undertaken to explore whether patient characteristics interact with treatment causing differences in mortality. It is concluded that patients on rtPA with a poorer prognosis achieve better long term survival than those with a good prognosis who may experience excess mortality on rtPA.

8.1 Introduction

Recombinant tissue plasminogen activator (rtPA) given as early as possible (< 6 hours) after acute ischaemic stroke favours good functional outcome (mRS 0 to 2 versus 3 to 6) with an increase in absolute benefit of 42 (95% CI 19 to 66) per 1000 treated (Wardlaw et al., 2012a). Despite this benefit rtPA is known to increase the risk of early symptomatic intracranial haemorrhage (SICH) by 58 (95% CI 49 to 68) per 1000. This is the most likely explanation of a known increase in early mortality (≤7 days) with around 25 more deaths (95% CI: 11 to 39) for every thousand patients treated with rtPA. On excluding those deaths that could be attributed to haemorrhage the difference in deaths between placebo and rtPA was no longer statistically significant with a P-value of 0.54. The majority of studies of rtPA in acute ischaemic stroke used a follow-up of three months or less (11 out of 12 included studies, Wardlaw et al., 2012). Little remains known about the long term impact of rtPA on patient mortality.

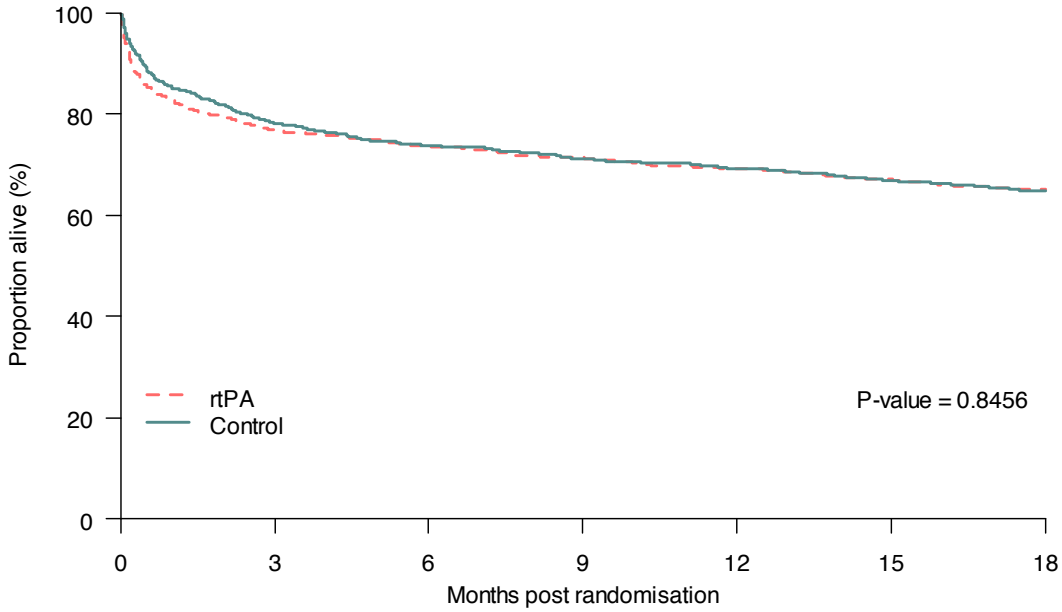
The third International Stroke Trial (IST-3) was introduced in the previous chapter were details of the design was provided. It was noted that the IST-3 had two follow-up phases. The first phase evaluated the harms and benefits of rtPA in acute

ischaemic stroke patients treated within six hours of stroke onset (The IST-3 collaborative group, 2012a). The second phase explored the impact of rtPA on long term survival (The IST-3 collaborative group, 2013). Details of the IST-3 collaborative group's analysis of the long term follow-up data are now discussed.

Ten out of the twelve contributing countries followed all patients recruited to their centres for a total of 18 months after randomisation. This comprised just over 77% of the original sample size (2348/3035). Those contributing countries were: Austria, Belgium, Canada, Italy, Mexico, Poland, the United Kingdom, Australia, Norway, and Sweden. Eighteen month mortality rates were similar between the two treatment arms with 34.9% and 35.1% on rtPA and control respectively. The IST-3 collaborative group formally compared the survival distributions under treatment and control using a log-rank test declaring no statistically significant difference between the survival rates under treatment with rtPA and control (P-value = 0.8456). However, the use of the log-rank test is not strictly appropriate for these data given the clear violation of the proportional hazards assumption (Figure 8-1) attributable to an early hazard of death in the rtPA treated arm. A test of the proportional hazards assumption on the same data, i.e., the 2348 patients, resulted in its rejection with a P-value of 0.0061 (Grambsch and Therneau, 1994).

It seems self-evident that for the proportion of those dead by 18 months to be approximately equal between treatment arms (34.9% and 35.1%) that some patients must experience an excess of mortality from treatment whilst others a survival benefit.

There are two patient characteristics that are plausibly related with survival: (i) the severity of the initial stroke; and (ii) the delay between stroke onset and treatment. The aim of this chapter is to explore these features using data from the IST-3: one of the largest trials of rtPA carried out to date in acute ischaemic stroke patients with also the longest record of patient follow-up.



rtPA	1169	898	859	824	795	769	746
Control	1179	920	869	832	806	779	753

Figure 8-1 Kaplan-Meier survival curves reproduced as per the data used in the eighteen month follow-up paper.

8.2 Survival analysis methods

In a standard clinical trial - cross-over trials excluded - patients are randomised to receive one of *k* possible treatments and followed for a fixed period of time recording the occurrence of any events of interest. Analyses which formally incorporate the recorded *time-to-event* data are termed *survival analyses*. The standard vernacular denotes events (e.g., death) as *failures* occurring at some observed time, or *failure time*. A subset of those recruited to IST-3 was followed for 18 months for the occurrence of all-cause mortality. Of course not all patients died during follow-up meaning that a proportion of the patients had no observed failure times: such data must be handled appropriately. In what follows, a brief exploration of the survival methods adopted throughout this chapter is provided.

8.2.1 Failure times, censoring times and the survival function

Patients who meet the trial inclusion criteria enter the trial as and when they become available. Various ethical and monetary constraints mean that typically the trial design will not be an 'observe all till failure' approach with instead patients being monitored via scheduled follow-up appointments for a fixed period of time. Those patients alive at the end of the 18 month follow-up in the IST-3 trial were known to have survived for at least 18 months. Such patients can be regarded as *right-censored* were the event time is unobserved and therefore censored by virtue of trial administration (e.g., Patient 4 in Figure 8-2 who in theory could fail at any point after 18 months). Similarly, patients who drop out of the trial during follow-up can still provide some information since it will be known that they survived up until the point at which they dropped out, therefore, such patients are also *right-censored* (e.g., Patient 1 in Figure 8-2).

The original analysis of long-term follow-up in the IST-3 was restricted to those patients with planned eighteen month follow-up (The IST-3 collaborative group, 2013). The analysis presented in this chapter adopts a more dynamic censoring scheme which attempts to incorporate all patients. In Portugal and Switzerland, planned follow-up was only for six months. Therefore any patient with a returned six month form but no record of death was censored at 183 days. If the six month form was unavailable but the date last known to be alive was greater than six months then the patient was censored at 183 days, or if this was less than six months then the last known date to be alive was used. There were nine Portuguese patients with eighteen month forms; 18 month follow-up was not planned for these patients so they were censored at 183 days. For those patients randomised in: Austria, Belgium, Canada, Italy, Mexico, Poland, the United Kingdom, Australia, Norway, and Sweden the following censoring scheme was applied. If the six month form was available and the eighteen month form was unavailable and there was no date when they were last known to be alive, the patient was censored at 183 days. If the six month form was available *or* was unavailable, the eighteen month form was available and there was no date last known to be alive then the patient was censored at 548 days. If neither the six month nor the eighteen month forms were available and the date last known

to be alive was less than eighteen months then the patient was censored at this date, if it exceeded eighteen months then the patient was censored at 548 days. One Portuguese patient was excluded from all analyses since neither of the forms was available nor was there a date last known to be alive.

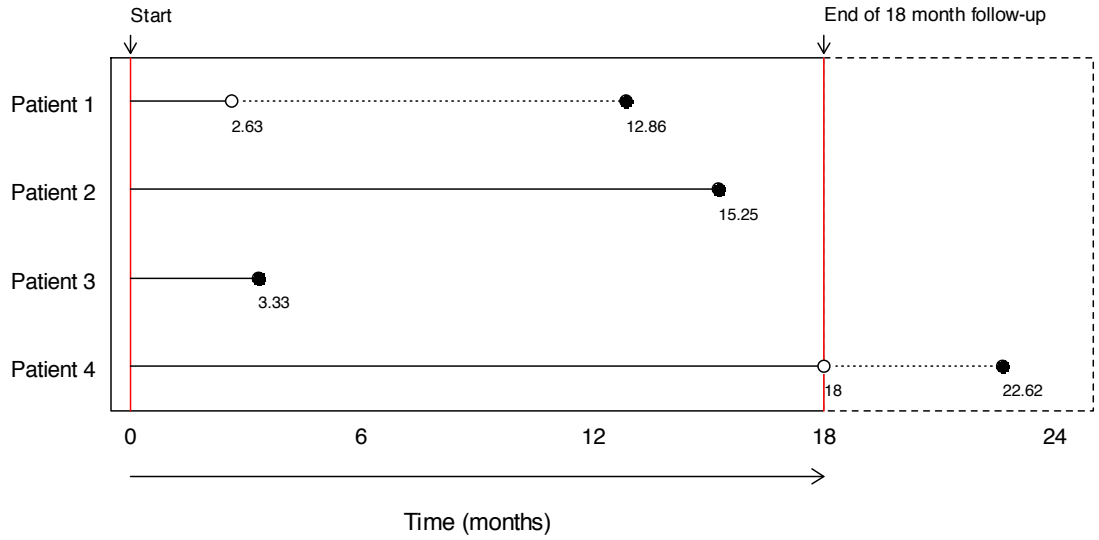


Figure 8-2 Time profiles for four hypothetical patients observed for 18 months for the occurrence of morality. Solid black circles denote failure times whilst open circles denote censoring times. Dotted lines illustrate potential *unobserved* follow-up.

At this stage two random variables can now be defined using more general terminology: patient survival time (T_S) and patient censoring time (T_C). An assumption of any formal statistical analysis of survival data is that these two random variables remain independent of one another, i.e., $T_S \perp T_C$, meaning that the underlying reason for any censoring should not depend on the outcome itself. The observed time is the minimum of the two, i.e., $T = \min(T_S, T_C)$. Each patient therefore provides an observed time and an event type (t_i, δ_i) where δ_i is an indicator function equal to 1 if $t_i \in T_S$ or 0 if $t_i \in T_C$. A cumulative density function is defined such that survival can be described as the probability of surviving beyond some time point, t,

$$S(t) = P(T_S \geq t) = 1 - F_S(t) \quad \forall\, t > 0\,, \tag{8.1}$$

where T_S is as previously defined and $F_S(t) = P(T_S < t)$ is the cumulative density function (Collett, 2003). The survival function is a non-increasing function equal to 1 at time point 0 decreasing to 0 as time extends to infinity. A similar function can be described for T_C the censoring function $C(t)$ (van Houwelingen and Putter, 2012).

8.2.2 The Kaplan-Meier estimator

There are two ways in which the survival function can be estimated. First, it is possible to make some parametric assumption and fit a known distribution which is defined on the real numbers, e.g., an exponential, Weibull, Gamma, log-normal distribution and so on. A drawback of this approach is a lack of flexibility, for example consider the exponential distribution which is fully specified by a single parameter; this means that a single constant hazard must hold which is unlikely to be true of, or applicable to, most clinical settings. Second, a non-parametric approach may be adopted making no distributional assumptions. The most frequently used is the Kaplan-Meier (*KM*) estimator (Kaplan and Meier, 1958). This was used earlier on in this chapter without specification (Figure 8-1); some theory is now provided.

The *KM* estimate provides a non-parametric estimate of the probability of failure at any given time point,

$$\hat{S}_{KM}(t) = \prod_{j=1}^{i} \frac{s_j}{r_j}, \tag{8.2}$$

for $t_{(i)} \le t < t_{(i+1)}, i = 1,\ldots,m$ with m event times. An estimate of the standard error can be obtained by first recognising that the *KM* estimate can be described as the product of a series of binomial estimates of conditional probabilities, $\hat{p}_j$, with s_j successes out of r_j trials such that $s_j \sim Bi(r_j, p_j)$. A simple rearrangement of the asymptotic property that $V(\log X) \approx V(X)/E(X)^2$ results in the following estimate (Greenwood, 1926):

$$SE\left(\hat{S}_{KM}(t)\right) = \hat{S}_{KM}(t)\sqrt{\sum_{j=1}^{i} \frac{r_j - s_j}{r_j s_j}}. \tag{8.3}$$

This can be used to obtain $(1 - \alpha)$ 100% point-wise confidence intervals for the *KM* estimate at each specified time point, i.e., $\hat{S}_{KM}(t) \pm z_{1-\alpha/2} SE\left(\hat{S}_{KM}(t)\right)$, where $z_{1-\alpha/2}$ is obtained from the standard Normal distribution. This can sometimes produce invalid intervals that fall out with the 0 and 1 boundary limits. This can be handled by adopting a transformation of the KM estimator. Some common transformations include the log and the log(- log) (van Houwelingen and Putter, 2012).

8.2.3 The hazard function and the modelling of covariates

The hazard function describes the instantaneous rate of failure. It is the limit of the probability that an individual experiences the event of interest in the next small time period of observation (Δt) as Δt tends to zero *given* that they survived up until time t,

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t}, \qquad \textbf{(8.4)}$$

in this way the hazard function incorporates the ageing process (Collett, 2003). It should be noted that the hazard and survival functions are both related. Appealing to standard probability theory equation (8.4) can be expressed as follows:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P\left(t \le T < t + \Delta t\right) / P\left(T \ge t\right)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \left\{ \frac{F(t + \Delta t) - F(t)}{\Delta t} \right\} \cdot \frac{1}{S(t)} \\ &= \frac{f(t)}{S(t)} \end{aligned}$$

Since $S(t)$ is the survival function as defined previously (equation (8.1)), and $f(t)$ is the derivative of $F(t)$, in the limit, as Δt approaches zero (Collett, 2003). By integrating $h(t)$ over time, t, the cumulative hazard rate $H(t)$ is obtained for which the survival function can be expressed as $S(t) = \exp(-H(t))$. This means that the focus of the analysis can easily be translated from hazard to survival and *vice versa*.

The most popular method for incorporating covariates into a survival analysis is the *proportional hazards model* (Harrell, 2001). In this way it is only necessary to specify a regression model through the hazard function, not the probability density function (Collett, 2003):

$$h(t \mid \boldsymbol{z}) = h_0(t)\exp\left(\boldsymbol{\beta}^T \boldsymbol{z}\right). \qquad \textbf{(8.5)}$$

In general, the baseline hazard, denoted by $h_0(t)$, expresses how the hazard function, $h(t \mid \boldsymbol{z})$, varies with survival time whilst the relative hazard function, $\exp(\boldsymbol{\beta}^T \boldsymbol{z})$, determines how the hazard changes with $\boldsymbol{z}$, the observed patient characteristics. Note that the PHM is the only model that will be considered for the analysis of survival data in this thesis.

The only requirement now is to suitably handle $h_0(t)$ which can be done by either adopting some parametric specification (e.g., exponential or Weibull etc.) or else by leaving it completely unspecified under the Cox Proportional Hazards model (Cox PHM). Before exploring the Cox model further it is important to discuss some assumptions. As stated in equation (8.5) the Proportional Hazards Model (PHM) has three identifiable assumptions that must hold in order that the resulting parameter estimates may be correctly interpreted (Harrell et al., 1988): (i) all continuous measurements have a particular shape with respect to the log hazard of the outcome (the simplest being a linear association, although a more complicated non-linear association may be pursued using splines (Harrell, 2001) or fractional polynomials (Sauerbrei and Royston, 1999)); (ii) the individual effects act additively with respect to the log hazard of the outcome; and (iii) the hazards are assumed to be proportional. The first two assumptions were covered in Chapter 2. The third assumption requires that the relative hazard function, $\exp(\boldsymbol{\beta}^T \boldsymbol{z})$, is constant with respect to time and therefore has the same effect on the hazard throughout follow-up. This means for instance that each predictor effect has the same relative effect no matter how far the patient is from the beginning or the end of the study. Consider two patients, one on treatment (z_1) and one on control (z_0). Using the PHM the interpretation of the effect of the treatment, β, is reduced to a single value in the following way,

$$\frac{h_2(t \mid z_1)}{h_1(t \mid z_0)} = \frac{h_0(t)\exp(\beta\, z_1)}{h_0(t)\exp(\beta\, z_0)} = \exp\left[\beta(z_1 - z_0)\right] = \exp(\beta)$$

with dummy variables coded 1 and 0 respectively for z_1 and z_0 the so called hazard ratio (HR) is determined as exp (β). This is the relative difference between the hazard of the individual on treatment and the hazard of the individual off treatment.

The Cox Proportional Hazards Model is by far the most commonly adopted method for regressing covariates on the hazard function which uniquely leaves the baseline hazard unspecified (Cox, 1972). The appeal of this approach is that no distributional assumptions need be made, yet $\boldsymbol{\beta}$, the vector of regression parameters associated with $\boldsymbol{z}$, can still be estimated so long as the proportional hazards assumption holds (Hosmer Jr et al., 2011). Cox proposed the *partial likelihood* function, an approach to estimation applicable when the full likelihood function can be expressed as the product of two separate functions specified on distinct parameters (e.g., in equation (8.5) the baseline hazard and the relative hazard function) of which only one is of interest (i.e., the estimation of $\boldsymbol{\beta}$) (Cole et al., 2014, Cox, 1975).

8.2.4 Handling non-proportional hazards

It was noted in the introduction of this chapter that an early hazard of excess deaths in rtPA treated patients caused a violation of the proportional hazards assumption, thus precluding the standard application of the Cox PHM. Formal tests of the proportional hazards assumption exist based on the linear correlation between the rank ordered failure times and scaled Schoenfeld's partial residuals (Grambsch and Therneau, 1994). Schoenfeld's residuals are calculated from the individual contribution that each observation makes to the derivative of the log partial likelihood (Hosmer Jr et al., 2011). It is useful to supplement such a test with a plot of time against the residuals (see Appendix A Figure 8-7). Three analytical approaches for handling non-proportional hazards are outlined below.

8.2.4.1 First approach: difference in estimated survival

The first and simplest approach considered stems from the idea that if you are interested in the difference between treatment arms then you should plot them (Senn,

2009, Senn et al., 1990). Empirical estimates of the differences in survival at specified time points can be obtained following a similar approach to that described by Frank E. Harrell, Jr. using his `survdiffplot` function – one of many useful functions from his library in R, `rms` (regression modeling strategies) which accompanies his book by the same name (Harrell, 2001, Harrell, 2013). There is no requirement for the Kaplan-Meier estimate to assume proportional hazards. It is therefore possible to explore the differences between the empirical estimates of survival probabilities on control and on rtPA at any observed time point, *t*.

Let $\hat{D}(t_i)$ denote the difference between the Kaplan-Meier estimates of survival, control minus rtPA, at the i^{th} time, t_i,

$$\hat{D}(t_i) = \hat{S}_{KM}^{control}(t_i) - \hat{S}_{KM}^{rtPA}(t_i) \,. \tag{8.6}$$

The standard error of the difference can be estimated from the individual standard errors for the Kaplan-Meier estimates of survival on control and on rtPA,

$$SE\left(\hat{D}(t_i)\right) = \sqrt{SE\left(\hat{S}_{KM}^{control}(t_i)\right)^2 + SE\left(\hat{S}_{KM}^{rtPA}(t_i)\right)^2} \,, \tag{8.7}$$

where the associated Kaplan-Meier standard errors are obtained using equation (8.3).

It logically follows that in order to test whether the absolute effect of treatment defined by equation (8.6) differs by some prognostic factor then the *difference of the differences* may be compared using estimates obtained in an identical manner to equations (8.6) and (8.7). The covariates of primary interest are: delay to randomisation; and patient's predicted prognosis. The motivation is to determine whether or not there is any significant long term or short term harm from rtPA relative to control after adjusting for these identifiable baseline measurements.

8.2.4.2 Second approach: restricted time interval Cox PH models

Whilst the method described previously is appealing in its simplicity, it is not possible to make efficient adjustments for multiple patient characteristics. In addition all the limitations associated with the categorisation of continuous predictors are imposed (Altman and Royston, 2006).

From a statistical perspective the most direct way to handle the non-proportional hazards problem would be to include an interaction between treatment and time to event thus fully accounting for the fact that the effect of rtPA is non-constant. Putter et al give an excellent illustration as to how this can be presented (Putter et al., 2005). However in the presence of treatment interactions with patients baseline characteristics it is less clear how the findings can be easily presented to medical audiences. To overcome this, it is possible to split time into distinct windows or epochs of observation and describe the effect of treatment in the presence of treatment interactions with patients baseline characteristics (Harrell, 2001). It should be stressed though that when splitting time in this manner the estimated treatment effects in the later epochs are dependent on the survival of the preceding time periods.

8.2.4.3 Third approach: using binary logistic regression

The most straightforward analysis would be to ignore the available time to event data and instead restrict the analysis of mortality to those with planned 18 month follow up using a simple multivariable binary logistic regression. Tests for interactions between treatment and patient characteristics then easily follow.

8.3 Results

The third International Stroke Trial (IST-3) recruited 3035 patients between May 2000 and July 2011. The original findings from the long term follow-up published by the IST-3 Collaborative group used 2348 (77.4% of 3035) of the original sample of patients for which 18 month follow-up had been planned (The IST-3 collaborative group, 2013). The censoring scheme applied in this chapter excludes only one patient from the rtPA arm. At the end of the 18 month follow-up period, 34.9% (529/1514) of those on rtPA and 35.5% (539/1520) of those on control were dead. The baseline characteristics of those randomised to each arm are summarised in Table 8-1. Patient characteristics in the IST-3 were described in detail in the previous chapter; as expected good balance was achieved between the two arms due to a minimisation algorithm employed in randomisation.

Two patients in the control group did not meet the inclusion criteria. This was highlighted in the initial publication (The IST-3 collaborative group, 2012a) with the following note:

> Two patients in the control group were randomly assigned at more than 6 h (protocol violation). One of these was recorded as having severe swelling on the randomisation scan, because the stroke had in fact occurred about 24 h earlier.

These two patients (enrolled at 10.83 and 24 hours) were retained in the following analysis, although excluded in a per protocol sensitivity analysis (see Appendix B).

Again a test of proportional hazards (Grambsch and Therneau, 1994) suggests that this does not hold true for treatment in IST-3 (N = 3034) with a P-value of 0.0011 (NB this is slightly different from the test presented at the beginning of this chapter which was based on those 2348 patients with planned 18 month follow-up). Inspection of the Scaled Schoenfeld residuals (see Appendix A Figure 8-7) suggests a decreasing treatment effect in the first two weeks of follow-up which was largely stable thereafter.

Table 8-1 Baseline characteristics of IST-3 patients included in 18-month follow-up.

Baseline variables collected before randomisation	rtPA (1514)	Control (1520)
Age (Median, IQR)	81 (72 to 86)	81 (71 to 86)
NIHSS (Median, IQR)	11 (6 to 18)	11 (6 to 17)
Time to randomisation (Median, IQR)	3.9 (2.9 to 4.9)	3.9 (2.9 to 4.8)
Sex, female (n, %)	781 (52)	788 (52)
Stroke Syndrome (n, %)		
TACI	639 (42)	666 (44)
PACI	595 (39)	551 (36)
LACI	168 (11)	164 (11)
POCI	110 (7)	136 (9)
Other	2 (<1)	3 (<1)
Country (n, %)		
Northwest Europe	792 (52)	797 (52)
Scandinavia	251 (17)	250 (16)
Australasia	89 (6)	90 (6)
Southern Europe	203 (13)	204 (13)
Eastern Europe	174 (11)	173 (11)
Americas	5 (<1)	6 (<1)
Expert reader's assessment of acute ischaemic change[1] (n, %)		
Scan completely normal	140 (9)	129 (9)
Abnormal scan, no sign of acute ischaemic change	742 (49)	781 (52)
Signs of acute ischaemic change	624 (41)	600 (40)

Abbreviations: IQR – Inter Quartile Range; NIHSS – National Institute of Health Stroke Scale; rtPA - recombinant tissue plasminogen activator; LACI - lacunar circulation infarcts; PACI - partial anterior circulation infarcts; POCI - posterior circulation infarcts; and TACI - total anterior circulation infarcts.
1 – rtPA with 8 missing and control 10 missing

8.3.1 First approach: absolute difference in survival

The difference of survival functions (control minus rtPA) is summarised in the plot below (Figure 8-3). Supposing for a moment that proportional hazards had in fact held (i.e., no excess mortality in the first few weeks of treatment) and that rtPA had no overall impact on mortality, in which case the plot of the difference in survival functions would simply be a random scatter about zero. However, this was not the case, which is immediately evident in the figure below.

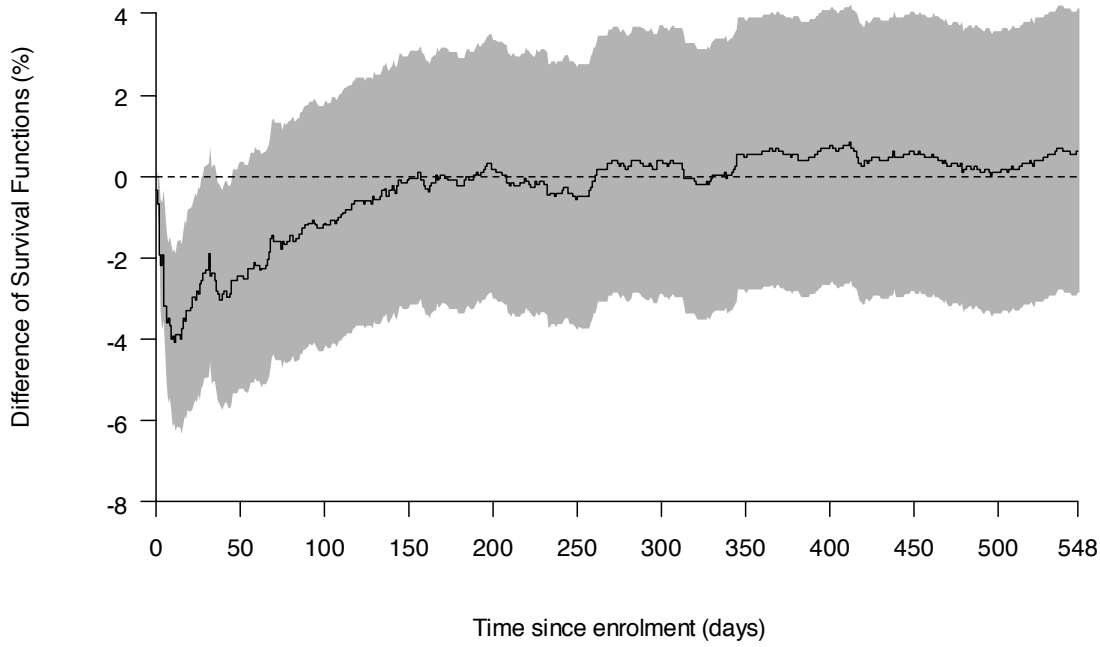


Figure 8-3 Difference of survival functions (%) amongst all patients, control minus rtPA (i.e., negative values indicate excess mortality in rtPA). Grey band highlights 95% point-wise CI.

In particular, by seven days those treated with rtPA experienced an excess of mortality over control of -3.46% (95% point-wise CI: -5.44 to -1.49), there was no excess mortality in either groups by six months (183 days) -0.10% (95% point-wise CI: -3.25 to 3.06) nor by 18 months (548 days) 0.63% (95% point-wise CI: -2.86 to 4.13). See also Table 8-2.

It is again stressed that since the clear early hazard of death was not sustained in the long term and since the difference in survival functions coalesce on zero (Figure 8-3)

then there must be groups of patients for which there is significantly more harm and groups for which there is significantly more benefit.

8.3.1.1 Absolute difference in survival by delay in enrolment

Two categorisations of delay in enrolment were considered: <3 versus ≥3 hours; and <4.5 versus ≥4.5 hours. The point estimates for strata specific differences in mortality between control and rtPA are provided in Table 8-2 whilst a test of the size of these differences between strata for specific time points are provided in Table 8-3. Those treated less than three hours from enrolment appear to benefit from a reduction in long term mortality (548 days) which was significantly larger than the same difference amongst those treated 3 hours or later (0.09 with 95% CI: 0.01 to 0.17, P-value = 0.0317). Moving the dichotomy to a later cut-point of time (4.5 hours) yielded a non-significant difference in late mortality (0.03 with 95% CI: -0.04 to 0.10, P-value = 0.3751). A plot of the difference of the differences is provided in section 8.5 Appendix A on page 285, Figure 8-8 (A).

8.3.1.2 Absolute difference in survival by predicted prognosis

Again with reference to Table 8-2 and Table 8-3 there was an indication that those with a good prognosis (<50%) experienced an excess of long term mortality (548 days) when treated with rtPA over control of 4.32% (95% point-wise CI: 8.48 to 0.17). There was evidence that this was significantly different from the size of the effect experienced by those with a poor predicted prognosis (≥50%) (-0.08 with 95% CI: -0.14 to -0.02, P-value = 0.0091). This effect was similar for a cut-point of 40% (-0.08 with 95% CI: -0.14 to -0.02, P-value = 0.0090) but not for 20% (-0.02 with 95% CI: -0.09 to 0.05, P-value = 0.5178) (see Table 8-2). A plot of the difference of the differences is provided in section 8.5 Appendix A on page 285, Figure 8-8 (B).

Table 8-2 Kaplan Meier estimates of mortality with 95% Point-wise CIs post stroke, with absolute difference of control minus rtPA.

	Control (%)		rtPA (%)		Difference (control - rtPA)	
Days since enrolment	**Deaths (n)**	**KM estimate (95% CI)**	**Deaths (n)**	**KM estimate (95% CI)**	**Estimate (95% CI)**	**P-value**
All patients, (Number of patients control versus rtPA, 1520 vs. 1514)						
7	102	6.71 (5.44 to 7.96)	154	10.17 (8.64 to 11.68)	-3.46 (-5.44 to -1.49)	0.0006
183	407	26.80 (24.54 to 29.00)	407	26.90 (24.63 to 29.10)	-0.10 (-3.25 to 3.06)	0.9510
548	539	36.56 (34.03 to 38.99)	529	35.93 (33.41 to 38.35)	0.63 (-2.86 to 4.13)	0.7235
Delay to randomisation						
Time to randomisation, < 3 hours, (*395 vs. 406*)						
7	35	8.86 (6.01 to 11.62)	41	10.10 (7.12 to 12.98)	-1.24 (-5.29 to -2.82)	0.5497
183	142	35.95 (31.04 to 40.51)	130	32.02 (27.33 to 36.41)	3.93 (-2.63 to 10.49)	0.2401
548	184	47.84 (42.55 to 52.65)	161	40.62 (35.54 to 45.30)	7.22 (0.21 to 14.23)	0.0434
Time to randomisation, ≥ 3 hours, (*1125 vs. 1108*)						
7	67	5.96 (4.56 to 7.33)	113	10.20 (8.40 to 11.96)	-4.24 (-6.50 to -1.99)	0.0002
183	265	23.59 (21.06 to 26.03)	277	25.03 (22.43 to 27.54)	-1.44 (-5.00 to 2.12)	0.4287
548	355	32.58 (29.72 to 35.33)	368	34.20 (31.29 to 37.00)	-1.62 (-5.62 to 2.38)	0.4270
Time to randomisation, < 4.5 hours, (*994 vs. 983*)						
7	74	7.44 (5.80 to 9.06)	110	11.19 (9.20 to 13.14)	-3.75 (-6.30 to -1.19)	0.0041
183	306	30.81 (27.88 to 33.62)	297	30.22 (27.29 to 33.03)	0.59 (-3.47 to 4.65)	0.7753
548	398	41.12 (37.91 to 44.16)	379	39.42 (36.23 to 42.45)	1.70 (-2.70 to 6.11)	0.4489
Time to randomisation, ≥ 4.5 hours, (*526 vs. 531*)						
7	28	5.32 (3.39 to 7.22)	44	8.29 (5.91 to 10.60)	-2.96 (-5.99 to 0.07)	0.0552
183	101	19.23 (15.79 to 22.53)	110	20.76 (17.23 to 24.13)	-1.53 (-6.35 to 3.30)	0.5353
548	141	27.89 (23.84 to 31.72)	150	29.41 (25.31 to 33.28)	-1.52 (-7.12 to 4.08)	0.5941

Table 8-2 Continued from previous page

	Control (%)		rtPA (%)		Difference (control - rtPA)	
Days since enrolment	**Deaths (n)**	**KM estimate (95% CI)**	**Deaths (n)**	**KM estimate (95% CI)**	**Estimate (95% CI)**	**P-value**
Subgroups of differing stroke severity						
Predicted prognosis < 50%, (*538 vs. 520*)						
7	8	1.49 (0.46 to 2.50)	14	2.69 (1.29 to 4.07)	-1.21 (-2.93 to 0.52)	0.1713
183	36	6.71 (4.57 to 8.81)	45	8.67 (6.22 to 11.06)	-1.96 (-5.18 to 1.26)	0.2323
548	55	10.71 (7.99 to 13.36)	74	15.03 (11.80 to 18.15)	-4.32 (-8.48 to -0.17)	0.0415
Predicted prognosis ≥ 50%, (*982 vs. 994*)						
7	94	9.57 (7.71 to 11.39)	140	14.08 (11.89 to 16.22)	-4.51 (-7.35 to -1.67)	0.0018
183	371	37.78 (34.67 to 40.74)	362	36.42 (33.36 to 39.35)	1.36 (-2.90 to 5.62)	0.5330
548	484	50.60 (47.30 to 53.70)	455	46.78 (43.52 to 49.85)	3.82 (-0.67 to 8.32)	0.0956
Predicted prognosis < 40%, (*378 vs. 351*)						
7	6	1.59 (0.32 to 2.84)	8	2.28 (0.71 to 3.83)	-0.69 (-2.70 to 1.31)	0.4991
183	17	4.51 (2.39 to 6.58)	27	7.72 (4.88 to 10.47)	-3.21 (-6.70 to 0.28)	0.0715
548	27	7.53 (4.75 to 10.23)	42	12.56 (8.92 to 16.05)	-5.03 (-9.53 to -0.53)	0.0284
Predicted prognosis ≥ 40%, (*1142 vs. 1163*)						
7	96	8.41 (6.78 to 10.00)	146	12.55 (10.63 to 14.44)	-4.15 (-6.64 to -1.65)	0.0011
183	390	34.18 (31.37 to 36.87)	380	32.68 (29.93 to 35.32)	1.50 (-2.35 to 5.35)	0.4461
548	512	46.06 (43.02 to 48.94)	487	42.94 (39.96 to 45.77)	3.12 (-1.02 to 7.27)	0.1395
Predicted prognosis < 20%, (*99 vs. 107*)						
7	0	0.00 (0.00 to 0.00)	2	1.87 (0.00 to 4.40)	-1.87 (-4.44 to 0.70)	0.1534
183	1	1.01 (0.00 to 2.96)	3	2.80 (0.00 to 5.88)	-1.79 (-5.49 to 1.90)	0.3416
548	3	3.33 (0.00 to 6.97)	5	4.95 (0.61 to 9.10)	-1.63 (-7.26 to 4.01)	0.5719
Predicted prognosis ≥ 20%, (*1421 vs. 1407*)						
7	102	7.18 (5.83 to 8.51)	152	10.80 (9.17 to 12.41)	-3.63 (-5.73 to -1.52)	0.0007
183	406	28.60 (26.21 to 30.92)	404	28.73 (26.33 to 31.06)	-0.13 (-3.46 to 3.21)	0.9406
548	536	38.86 (36.21 to 41.40)	524	38.26 (35.62 to 40.80)	0.59 (-3.07 to 4.26)	0.7512

Table 8-3 Testing the difference of the differences.

Comparing groups	Estimate	SE.	P-value
Delay, <3 hours vs. ≥3 hours			
7	0.0301	0.0237	0.2043
183	0.0537	0.0381	0.1585
548	0.0884	0.0412	0.0317
Delay, <4.5 hours vs. ≥4.5 hours			
7	-0.0078	0.0202	0.6989
183	0.0212	0.0322	0.5104
548	0.0322	0.0364	0.3751
Prognosis, <50% vs. ≥50%			
7	0.0331	0.0170	0.0511
183	-0.0332	0.0272	0.2235
548	-0.0814	0.0312	0.0091
Prognosis, <40% vs. ≥40%			
7	0.0346	0.0163	0.0343
183	-0.0471	0.0265	0.0759
548	-0.0815	0.0312	0.0090
Prognosis, <20% vs. ≥20%			
7	0.0176	0.0169	0.2998
183	-0.0167	0.0254	0.5117
548	-0.0222	0.0343	0.5178

8.3.2 Second approach: Cox PHMs in distinct epochs of time

As previously noted there are two main flaws in adopting an approach which looks at the difference in survival functions: first, only one interaction can be assessed at a time; and second, the two strictly continuous baseline measurements are handled in a rather inefficient way through necessary categorisation. Under this approach, such an evaluation of the effect of treatment is not as exhaustive as it should be. It was noted

earlier that the added complication of non-proportional hazards precludes the standard use of Cox PHMs fit with main effects (i.e., treatment, predicted prognosis and delay to treatment) and the two treatment interactions. In order to circumvent this problem whilst simultaneously incorporating the treatment interactions, the period of follow-up was split into epochs of time: 0 to 7 days; 7 to 90 days; 90 to 183 days; and 183 to 548 days. Within each follow-up period the effect of: treatment; prognosis; delay; treatment with delay interaction; and treatment with prognosis interaction was assessed using four separate Cox PHMs for each of which the proportional hazards assumption held with global test P-values > 0.3 (see legend of Figure 8-4).

Table 8-4 demonstrates the impact on the quality of the model fit for each period of observation assuming an increasingly larger model (i.e., adding interactions with treatment) as well as reducing the fit from the full model (i.e., removing interactions with treatment). The Likelihood Ratio (LR) test was used to assess the individual nested model fits. The LR test is based on the chi-squared distribution and tests whether there is a significant contribution from the larger model to the amount of information that is already known through the smaller model (note that the *size* of the model relates to the number of parameters fit, see Chapter 2 for more detail). For early mortality (within at most 7 days) there was no additional information gained via the inclusion of interactions with treatment. A safe conclusion here may then be that the single main effects treatment HR (i.e., M_0) is the best estimate of the effect of rtPA on early (0 to 7 day) mortality, 1.54 (95%CI: 1.19 to 2.00). When building from the main effects fit in the interval 7 to 90 days there was a weak indication that treatment interacted with prognosis (χ^2=3.02) though not significantly. In the interval 90 to 183 days there was a strong suggestion that treatment interacted with delay (χ^2=4.86), though in the final epoch (183 to 548 days) the importance of predicted prognosis far out-weighed any interaction with delay (χ^2=7.37).

Figure 8-4 illustrates how the treatment hazard ratios for M_3 differ across the considered time intervals whilst interacting with delay and prognosis.

Table 8-4 Comparing Cox PHMs with and without treatment interactions. Note, n denotes the number of deaths, N the number at risk and df the degrees of freedom.

Model comparison	*n*	*N*	*LR* χ^2	df	P-value
Time interval: 0 to 7 days	251	3034			
M_1 vs M_0	-	-	1.40	1	0.2371
M_2 vs M_0	-	-	0.01	1	0.9386
M_3 vs M_0	-	-	1.42	2	0.4910
M_3 vs M_1	-	-	0.03	1	0.8743
M_3 vs M_2	-	-	1.14	1	0.2339
Time interval: 7 to 90 days	437	2783			
M_1 vs M_0	-	-	1.29	1	0.2554
M_2 vs M_0	-	-	3.02	1	0.0825
M_3 vs M_0	-	-	5.61	2	0.0604
M_3 vs M_1	-	-	4.32	1	0.0377
M_3 vs M_2	-	-	2.60	1	0.1070
Time interval: 90 to 183 days	126	2342			
M_1 vs M_0	-	-	4.86	1	0.0275
M_2 vs M_0	-	-	0.41	1	0.5247
M_3 vs M_0	-	-	4.86	2	0.0880
M_3 vs M_1	-	-	0.00	1	0.9524
M_3 vs M_2	-	-	4.46	1	0.0348
Time interval: 183 to 548 days	254	2213			
M_1 vs M_0	-	-	1.64	1	0.1994
M_2 vs M_0	-	-	7.37	1	0.0066
M_3 vs M_0	-	-	7.71	2	0.0212
M_3 vs M_1	-	-	6.10	1	0.0138
M_3 vs M_2	-	-	0.34	1	0.5622

Model Key

M_0	β_1*prognosis + β_2*delay + β_3*treatment
M_1	β_1*prognosis + β_2*delay + β_3*treatment + **β_4*delay:treatment**
M_2	β_1*prognosis + β_2*delay + β_3*treatment + **β_5* prognosis:treatment**
M_3	β_1*prognosis + β_2*delay + β_3*treatment + **β_4*delay:treatment + β_5* prognosis:treatment**

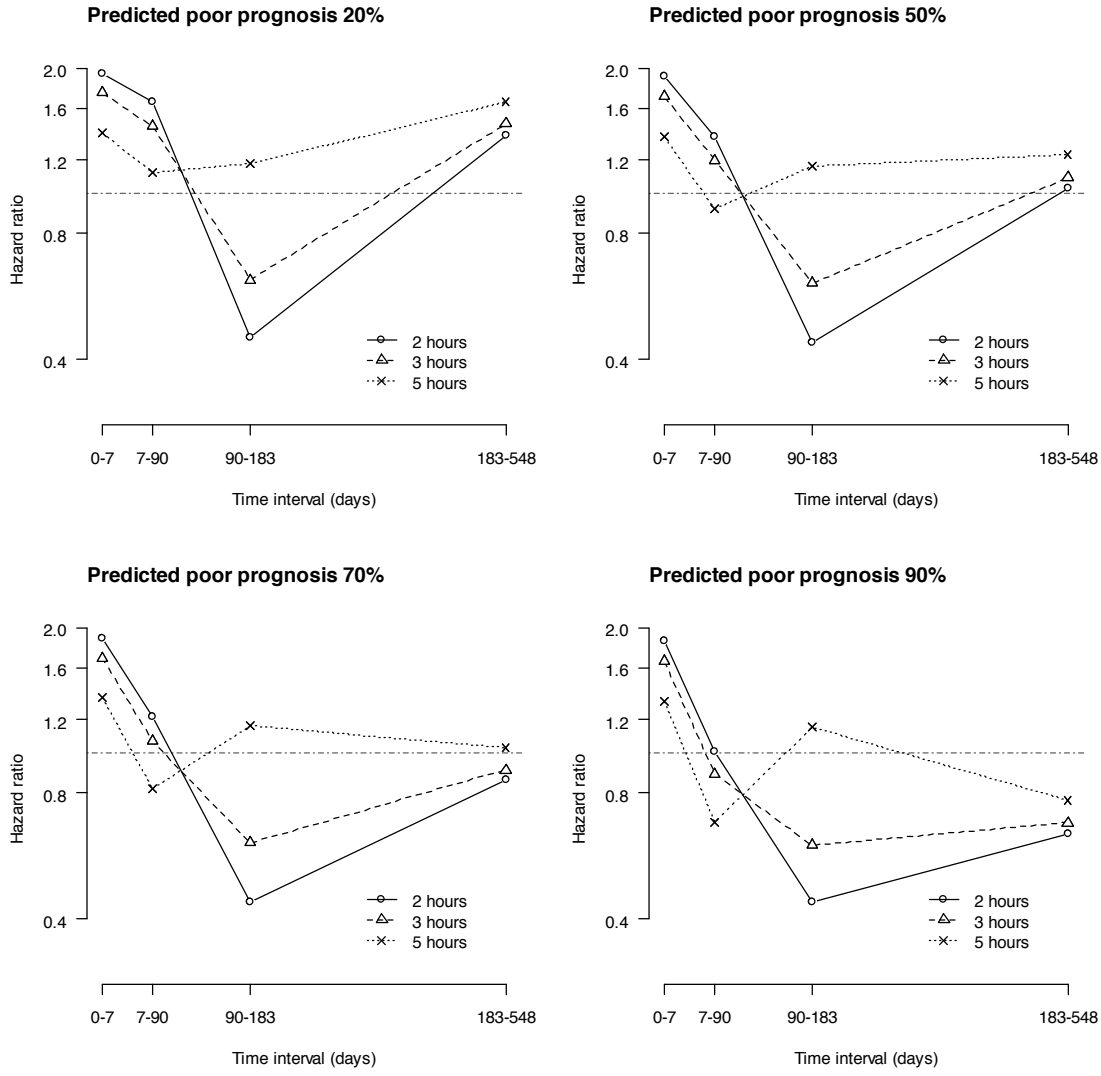


Figure 8-4 Treatment interaction hazard ratios per time epoch for distinct types of patients with fixed prognosis (20%, 50%, 70% and 90%) and fixed hours till enrolment (2, 3, and 5 hours). Note an HR>1 indicates an increase in the hazard of mortality and HR<1 a decrease. Global tests of proportional hazards for each Cox model fit within the four time windows, ordered with increasing time: 0.8040, 0.4480, 0.4360 and 0.3744.

8.3.3 Third approach: binary logistic regression analysis

The final approach discards all time to event information focusing on any deaths that occurred across the 18 month follow-up period for those patients on which 18 month follow-up was planned. Here the IST-3 data are restricted to the same data analysed in the Lancet paper, (The IST-3 collaborative group, 2013). A similar analysis was conducted in the cited publication, although no assessment was made for treatment interactions. This analysis fit a logistic regression adjusting for: age, NIHSS, and delay (all linear) and visible infarct on baseline scan yielding a treatment effect of 0.95 (95%CI: 0.78 to 1.16, P-value = 0.628).

In this secondary analysis, the same model comparison process as described in the previous section is followed only here binary logistic regression models are fitted instead of Cox PHMs. An interaction between delay and treatment (χ^2=0.46) was not as great as including an interaction between treatment and predicted prognosis (χ^2=7.78) and even less important in the presence of the treatment-prognosis interaction (χ^2=0.04) (Table 8-5). Figure 8-5 summarises the extent of the treatment effect on 18 month mortality and its interaction with patient prognosis and delay.

Table 8-5 Comparing Binary logistic regression model fits with and without treatment interactions. Note, n denotes the number of deaths, N the number at risk and df the degrees of freedom.

Model comparison	*n*	*N*	*LR* χ^2	df	P-value
18 month mortality	822	2348			
M_1 vs M_0	-	-	0.46	1	0.4990
M_2 vs M_0	-	-	7.78	1	0.0053
M_3 vs M_0	-	-	7.81	2	0.0201
M_3 vs M_1	-	-	7.36	1	0.0138
M_3 vs M_2	-	-	0.04	1	0.8402

Model Key

M_0	β_1*prognosis + β_2*delay + β_3*treatment
M_1	β_1*prognosis + β_2*delay + β_3*treatment + **β_4*delay:treatment**
M_2	β_1*prognosis + β_2*delay + β_3*treatment + **β_5* prognosis:treatment**
M_3	β_1*prognosis + β_2*delay + β_3*treatment + **β_4*delay:treatment + β_5* prognosis:treatment**

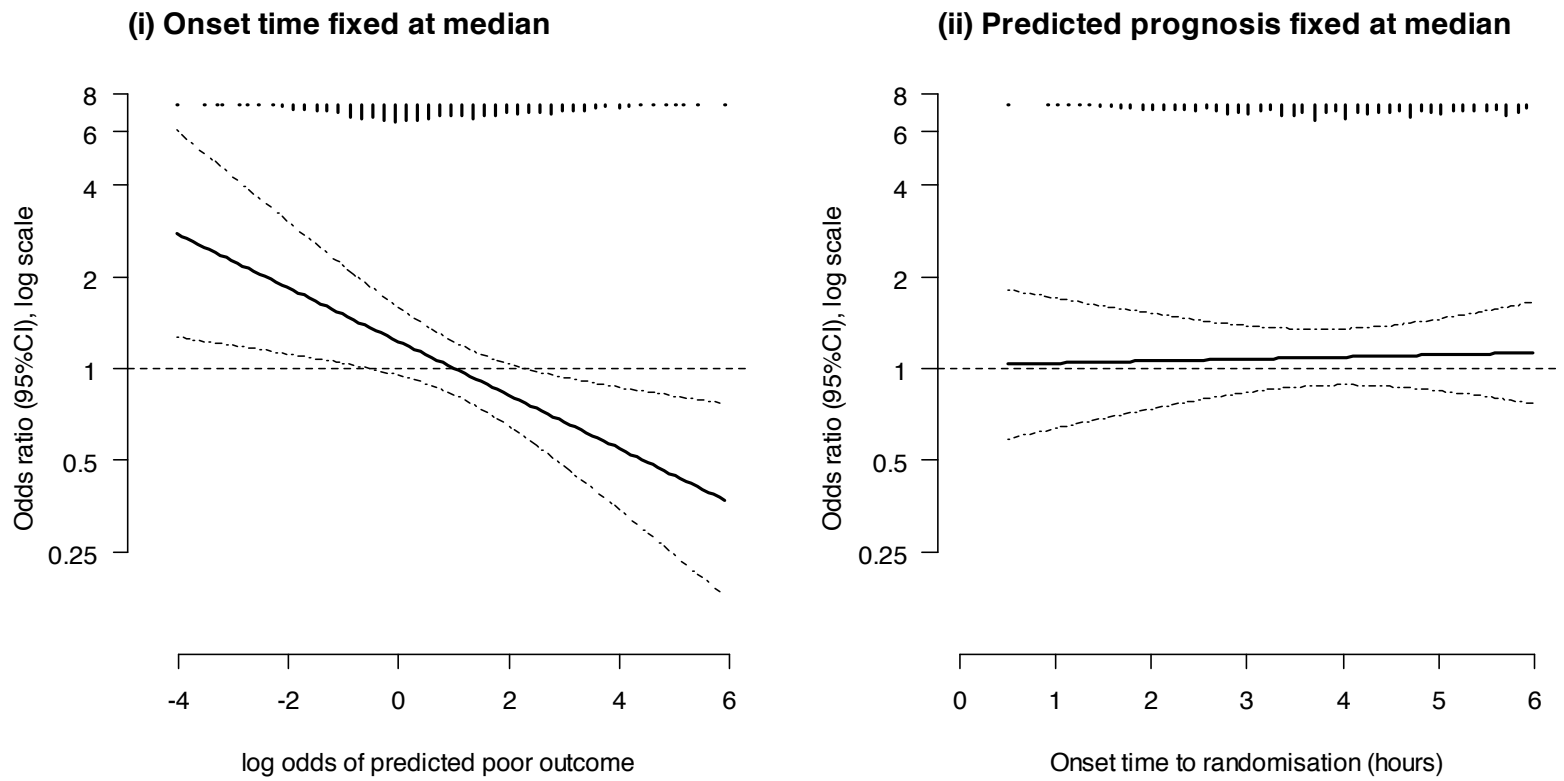


Figure 8-5 The effect of rtPA on 18 month mortality. (i) Treatment with rtPA OR across predicted prognosis (log odds scale) with onset time to randomisation fixed at median time (3.9 hours). The ORs for rtPA at the 25th, 50th and 75th percentiles are: 1.32 (95%CI 0.98 to 1.77); 1.09 (95%CI 0.88 to 1.34); and 0.83 (95%CI 0.66 to 1.04) respectively. (ii) Treatment with rtPA OR across onset time to randomisation with predicted prognosis fixed at median risk (64%, log odds = 0.57) at 1, 3 and 6 hours are: 1.04 (95%CI 0.64 to 1.70); 1.07 (95%CI 0.84 to 1.38); and 1.13 (95%CI 0.76 to 1.66) respectively. A global test of additivity (rtPA with prognosis and rtPA with delay) was significant with a P-value of 0.0208.

8.4 Discussion

This chapter has demonstrated that although rtPA does not correspond to any significant difference in mortality overall by 18 months, there is evidence to support the hypothesis that the effect of rtPA on mortality depends, to an extent, upon delay to randomisation and prognosis. When comparing the difference in Kaplan-Meier estimates of mortality between treatment arms across a binary categorisation of delay split at 3 hours it appeared that those treated within 3 hours experienced a beneficial absolute reduction in late mortality (548 days) compared to those randomised later, i.e., ≥3 hours (P-value = 0.0317). Those with a good predicted prognosis (risk of death or dependency <50%) fared better with regards to late mortality (548 days) on control than on rtPA when contrast to those with a poor predicted prognosis (P-value = 0.0091). However, in an adjusted analysis the relative improvement in the model fit when allowing treatment and delay to interact was not as great as when allowing treatment and prognosis to interact; this attenuation of the effect of delay may be explained by confounding. Reassuringly each of the approaches adopted to tackle the problem of non-proportional hazards reached a similar conclusion. In particular, that treatment with rtPA interacts with patient prognosis such that those with a good predicted prognosis may be subject to a higher risk of late mortality whilst those with a poor predicted prognosis experience a late survival benefit.

A recent IPDMA of nine completed trials (including IST-3) of recombinant tissue plasminogen activator (rtPA) in acute ischaemic stroke explored the impact of delay to randomisation upon the efficacy of rtPA on 90 day mortality (Emberson et al., 2014). The analysis presented by Emberson *et al*. found no evidence to support the hypothesis that the effect of rtPA on short term mortality interacts with delay to randomisation (P-value = 0.22); although they did note that a lack of statistical power to detect a true interaction likely played a part.

Three specific analytical approaches were adopted in this chapter in order to handle the problem of non-proportional hazards in the IST-3 data. It is important to reflect upon the pros and cons of each method so that recommendations may be made for future studies. The *difference of the differences* method has an appealing simplicity;

allowing inferences to be made from an easily obtained summary measure quantifying the differences in standard Kaplan-Meier estimates of survival for any given follow-up time. However, limitations include the need to categorise continuous variables and the inability to efficiently adjust for more than one characteristic at a time. In the context of the IST-3 data it is known that those randomised later where on average younger and had less severe strokes (Sandercock et al., 2011). Delay and prognosis should therefore not be interpreted separately as confounding plays an important role. This approach is best justified as a useful exploration tool but caution must be emphasised when drawing inferences.

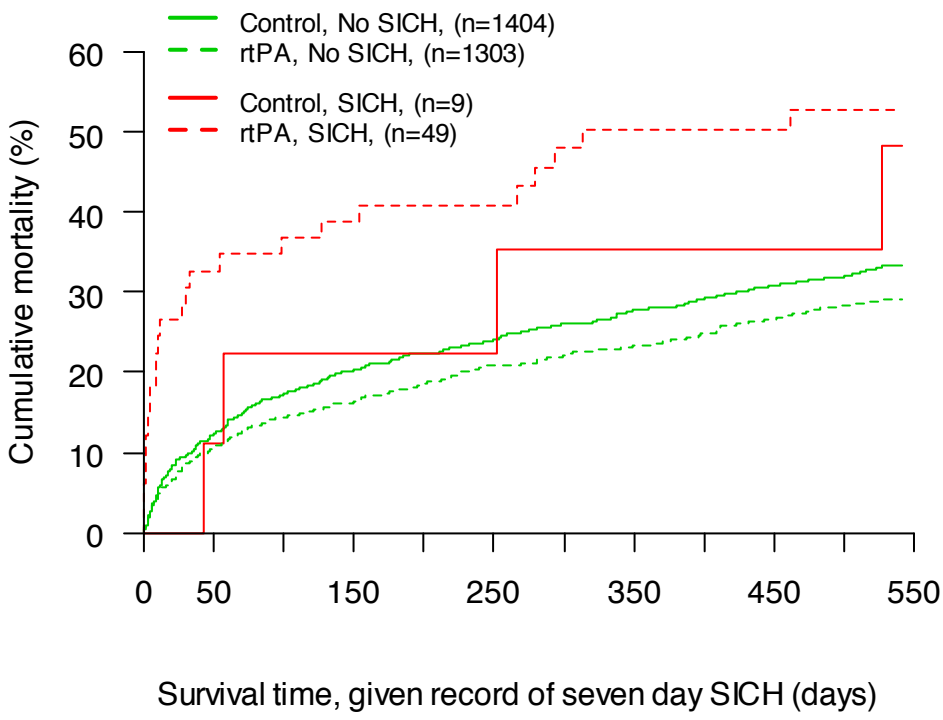
Fitting individual Cox PHMs within epochs of follow-up offers an interpretable way of describing the dependency of treatment upon time in the presence of non-proportional hazards. This is akin to the explicit inclusion of an interaction between treatment and follow-up time, i.e. a time-dependent effect; however, to also explore whether treatment – which already varies with time –varies with delay to randomisation and prognosis adds an extra layer of complexity which is difficult to express in a simple and clear way. Splitting follow-up time has a detrimental impact on the statistical power to detect true effects and is a penalty that must be weighed up with the appeal of this simpler summary.

Finally, a simple binary logistic regression model using only the data corresponding to those countries which followed patients up for 18 months was implemented. This is perhaps an oversimplification of the problem especially when a biological explanation for the early hazard of mortality is most likely attributed to an early risk of SICH (Figure 8-6). Nevertheless the logistic regression fit does reach the same conclusion that was drawn from a Cox PHM fit within the later time epoch (183 to 548 days) that predicted prognosis has an important interaction with the effect of rtPA on late mortality whilst delay to randomisation has little impact (see Figure 8-4 and Figure 8-5).

The conclusion of the previous chapter was that current prediction models for SICH post rtPA are similar to those for poor functional outcome. In general, they both separate patients into risk groups in which the highest risk group (from either event)

gain the largest absolute benefit from rtPA in reducing six month poor functional outcome. Therefore whilst early mortality is likely attributable to the increase in risk from SICH, it is not yet possible to reliably identify those that will likely suffer an SICH from those that will not as current predictions largely correlate with the predicted risk of poor functional outcome.

The direction of the results and the conclusions drawn were not affected by adopting a per protocol analysis (see Appendix B).



No SICH	**7 days**	**183 days**	**541 days**
Control	49	305	431
rtPA	47	232	347
SICH			
Control	0	2	4
rtPA	9	20	25

Figure 8-6 Cumulative mortality plots of rtPA-treated versus control patients split by record of symptomatic intracranial haemorrhage (SICH).

8.5 Appendix A: additional plots

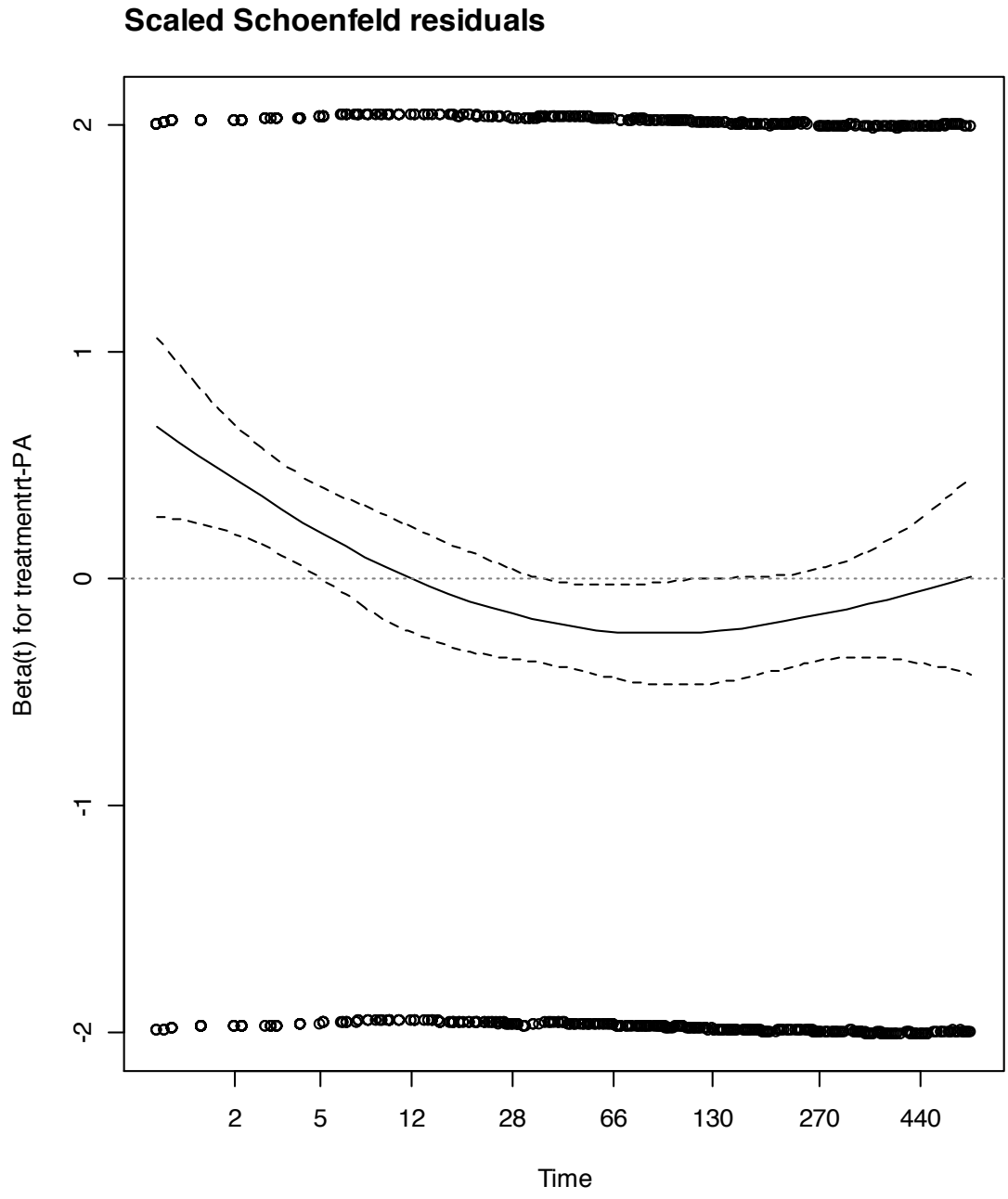
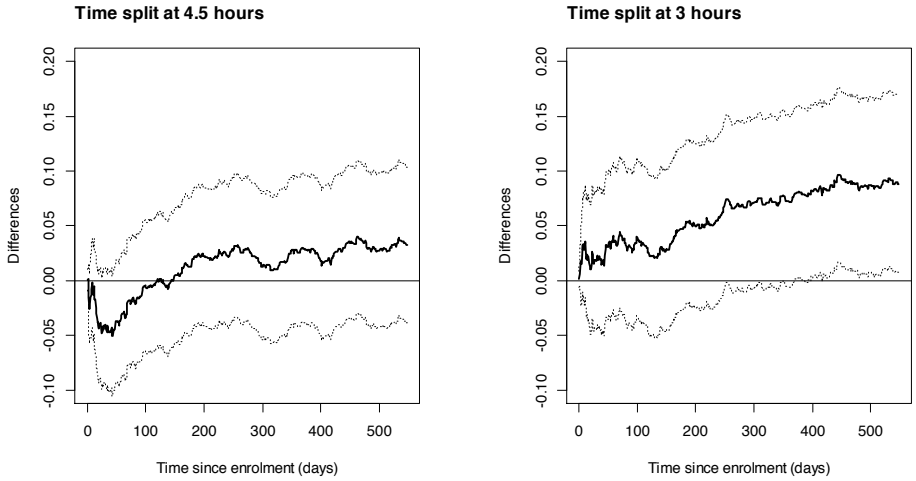
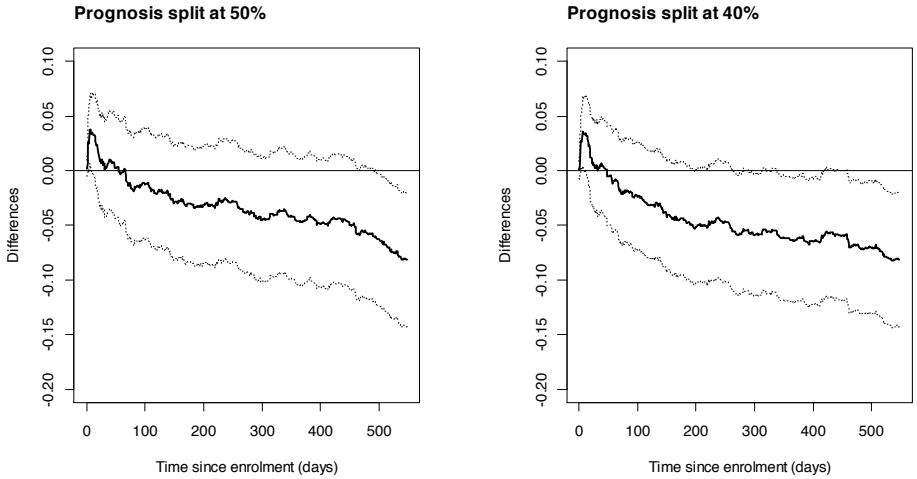


Figure 8-7 Scaled Schoenfeld residuals plotted against log time provide a visual assessment of the PH assumption. A formal test based on the Schoenfeld residuals gives a significant P-value of 0.0011.

A: Randomisation delay



B (i): Predicted prognosis



B (ii): Predicted prognosis

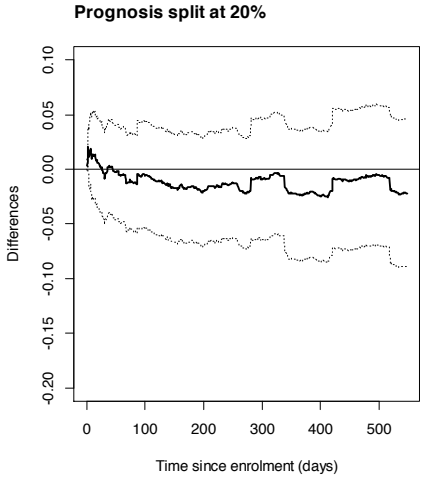


Figure 8-8 Plot of differences by various categorisations of: (A) randomisation time; and (B(i) and B(ii)) predicted prognosis.

8.6 Appendix B: per protocol sensitivity analysis

It is possible that the intention to treat analysis presented in this chapter could be different from a per protocol analysis. This was therefore conducted as a sensitivity analysis reducing the sample from 3034 patients to 2714. Here the analysis of the difference in mortality estimates based on the Kaplan-Meier estimates are repeated (Table 8-6 and Table 8-7). This did not lead to any qualitatively different conclusions with slight differences in estimates as would be expected.

Similarly, the process of comparing Cox PHMs fit to time epochs was replicated in these data (Table 8-8). Again, there was no qualitative difference in conclusions.

It appears safe to conclude that the findings presented in this chapter are robust irrespective of whether an intention to treat analysis or a per protocol analysis is followed.

Table 8-6 Per protocol analysis: Kaplan Meier estimates of mortality with 95% Point-wise CIs post stroke, with absolute difference of control minus rtPA.

	Control (%)		rtPA (%)		Difference (control - rtPA)	
Days since enrolment	**Deaths (n)**	**KM estimate (95% CI)**	**Deaths (n)**	**KM estimate (95% CI)**	**Estimate (95% CI)**	**P-value**
All patients, (Number of patients control versus rtPA, 1368 vs. 1346)						
7	84	6.14 (4.86 to 7.40)	125	9.29 (7.72 to 10.82)	-3.15 (-5.15 to -1.14)	0.0021
183	359	26.27 (23.90 to 28.57)	354	26.32 (23.93 to 28.64)	-0.05 (-3.37 to 3.26)	0.9762
548	479	36.23 (33.56 to 38.79)	464	35.47 (32.80 to 38.03)	0.76 (-2.93 to 4.45)	0.6868
Delay to randomisation						
Time to randomisation, < 3 hours, (*364 vs. 358*)						
7	33	9.07 (6.07 to 11.97)	32	8.94 (5.93 to 11.85)	0.13 (-4.05 to 4.30)	0.9523
183	131	35.99 (30.86 to 40.73)	113	31.56 (26.58 to 36.21)	4.42 (-2.47 to 11.32)	0.2082
548	170	48.04 (42.50 to 53.04)	137	39.17 (33.79 to 44.10)	8.87 (1.51 to 16.23)	0.0182
Time to randomisation, ≥ 3 hours, (*1004 vs. 988*)						
7	51	5.08 (3.71 to 6.43)	93	9.41 (7.57 to 11.22)	-4.33 (-6.60 to -2.06)	0.0002
183	228	22.75 (20.11 to 25.30)	241	24.42 (21.69 to 27.06)	-1.68 (-5.41 to 2.05)	0.3785
548	309	31.93 (28.90 to 34.82)	327	34.12 (31.03 to 37.07)	-2.19 (-6.42 to 2.03)	0.3093
Time to randomisation, < 4.5 hours, (*894 vs. 873*)						
7	62	6.94 (5.25 to 8.59)	91	10.42 (8.37 to 12.43)	-3.49 (-6.11 to -0.87)	0.0091
183	274	30.68 (27.59 to 33.64)	257	29.44 (26.35 to 32.40)	1.23 (-3.04 to 5.51)	0.5721
548	356	40.98 (37.59 to 44.19)	329	38.53 (35.16 to 41.73)	2.45 (-2.20 to 7.10)	0.3023
Time to randomisation, ≥ 4.5 hours, (*474 vs. 473*)						
7	22	4.64 (2.73 to 6.52)	34	7.19 (4.83 to 9.49)	-2.55 (-5.55 to 0.45)	0.0962
183	85	17.96 (14.43 to 21.35)	97	20.55 (16.82 to 24.12)	-2.59 (-7.62 to 2.43)	0.3120
548	123	27.21 (22.95 to 31.23)	135	29.77 (25.40 to 33.89)	-2.57 (-8.49 to 3.36)	0.3955

Table 8-6 Continued from previous page

	Control (%)		rtPA (%)		Difference (control - rtPA)	
Days since enrolment	**Deaths (n)**	**KM estimate (95% CI)**	**Deaths (n)**	**KM estimate (95% CI)**	**Estimate (95% CI)**	**P-value**
Subgroups of differing stroke severity						
Predicted prognosis < 50%, (*486 vs. 466*)						
7	6	1.23 (0.25 to 2.21)	13	2.79 (1.28 to 4.27)	-1.56 (-3.34 to 0.23)	0.0884
183	31	6.40 (4.20 to 8.56)	42	9.03 (6.39 to 11.60)	-2.63 (-6.03 to 0.76)	0.1287
548	48	10.41 (7.57 to 13.16)	68	15.42 (11.96 to 18.74)	-5.01 (-9.41 to -0.62)	0.0253
Predicted prognosis ≥ 50%, (*882 vs. 880*)						
7	78	8.84 (6.95 to 10.7)	112	12.73 (10.50 to 14.90)	-3.88 (-6.78 to -0.99)	0.0085
183	328	37.19 (33.92 to 40.3)	312	35.46 (32.22 to 38.55)	1.73 (-2.76 to 6.22)	0.4512
548	431	50.33 (46.82 to 53.6)	396	45.99 (42.53 to 49.24)	4.34 (-0.43 to 9.10)	0.0744
Predicted prognosis < 40%, (*345 vs. 317*)						
7	5	1.45 (0.18 to 2.70)	8	2.52 (0.78 to 4.24)	-1.07 (-3.21 to 1.06)	0.3247
183	16	4.65 (2.40 to 6.84)	26	8.23 (5.15 to 11.21)	-3.58 (-7.34 to 0.18)	0.0617
548	26	8.00 (4.99 to 10.91)	41	13.57 (9.60 to 17.36)	-5.57 (-10.45 to -0.7)	0.0252
Predicted prognosis ≥ 40%, (*1023 vs. 1163*)						
7	79	7.72 (6.07 to 9.34)	117	11.37 (9.41 to 13.29)	-3.65 (-6.19 to -1.11)	0.0048
183	343	33.56 (30.60 to 36.39)	328	31.88 (28.97 to 34.67)	1.68 (-2.38 to 5.74)	0.4184
548	453	45.64 (42.42 to 48.68)	423	42.18 (39.02 to 45.18)	3.46 (-0.93 to 7.85)	0.122
Predicted prognosis < 20%, (*96 vs. 100*)						
7	0	0.00 (0.00 to 0.00)	2	2.00 (0.00 to 4.71)	-2.00 (-4.74 to 0.74)	0.1531
183	1	1.04 (0.00 to 3.05)	3	3.00 (0.00 to 6.29)	-1.96 (-5.87 to 1.95)	0.3265
548	3	3.44 (0.00 to 7.20)	5	5.30 (0.65 to 9.72)	-1.86 (-7.79 to 4.08)	0.5403
Predicted prognosis ≥ 20%, (*1272 vs. 1246*)						
7	84	6.60 (5.23 to 7.96)	123	9.87 (8.20 to 11.51)	-3.27 (-5.41 to -1.12)	0.0028
183	358	28.18 (25.66 to 30.61)	351	28.19 (25.65 to 30.65)	-0.01 (-3.53 to 3.51)	0.9961
548	476	38.69 (35.88 to 41.38)	459	37.87 (35.06 to 40.56)	0.82 (-3.07 to 4.70)	0.6793

Table 8-7 Per protocol analysis: testing the difference of the differences

Comparing groups	Estimate	SE.	P-value
Delay, <3 hours vs. ≥3 hours			
7	0.0446	0.0243	0.0659
183	0.0610	0.0400	0.1270
548	0.1107	0.0433	0.0106
Delay, <4.5 hours vs. ≥4.5 hours			
7	-0.0094	0.0203	0.6433
183	0.0383	0.0337	0.2559
548	0.0502	0.0384	0.1917
Prognosis, <50% vs. ≥50%			
7	0.0233	0.0173	0.1795
183	-0.0436	0.0287	0.1292
548	-0.0935	0.0331	0.0047
Prognosis, <40% vs. ≥40%			
7	0.0257	0.0169	0.1285
183	-0.0526	0.0282	0.0624
548	-0.0903	0.0335	0.0070
Prognosis, <20% vs. ≥20%			
7	0.0127	0.0178	0.4756
183	-0.0195	0.0268	0.4676
548	-0.0267	0.0362	0.4601

Table 8-8 Per protocol analysis: comparing Cox PHMs with and without treatment interactions. Note, n denotes the number of deaths, N the number at risk and df the degrees of freedom.

Model comparison	*n*	*N*	*LR* χ^2	df	P-value
Time interval: 0 to 7 days	204	2714			
M_1 vs M_0	-	-	0.79	1	0.3729
M_2 vs M_0	-	-	0.13	1	0.7155
M_3 vs M_0	-	-	1.12	2	0.5705
M_3 vs M_1	-	-	0.33	1	0.5666
M_3 vs M_2	-	-	0.99	1	0.3198
Time interval: 7 to 90 days	395	2510			
M_1 vs M_0	-	-	1.14	1	0.2858
M_2 vs M_0	-	-	4.91	1	0.0267
M_3 vs M_0	-	-	7.70	2	0.0213
M_3 vs M_1	-	-	6.56	1	0.0104
M_3 vs M_2	-	-	2.79	1	0.0949
Time interval: 90 to 183 days	114	2111			
M_1 vs M_0	-	-	6.01	1	0.0143
M_2 vs M_0	-	-	1.10	1	0.3024
M_3 vs M_0	-	-	6.16	2	0.0460
M_3 vs M_1	-	-	0.15	1	0.6987
M_3 vs M_2	-	-	5.10	1	0.0240
Time interval: 183 to 548 days	230	1994			
M_1 vs M_0	-	-	2.69	1	0.1012
M_2 vs M_0	-	-	6.61	1	0.0102
M_3 vs M_0	-	-	7.57	2	0.0227
M_3 vs M_1	-	-	4.88	1	0.0272
M_3 vs M_2	-	-	0.96	1	0.3270

Model Key

M_0	β_1*prognosis + β_2*delay + β_3*treatment
M_1	β_1*prognosis + β_2*delay + β_3*treatment + **β_4*delay:treatment**
M_2	β_1*prognosis + β_2*delay + β_3*treatment + **β_5* prognosis:treatment**
M_3	β_1*prognosis + β_2*delay + β_3*treatment + **β_4*delay:treatment + β_5* prognosis:treatment**

Chapter 9: The impact of major haemorrhagic and arterial thrombotic events on mortality

Background and summary

Following an acute ischaemic stroke, patients may suffer an additional adverse event which can further complicate their recovery. In this chapter a large registry of stroke trial data is analysed assessing the impact that such events have upon patient mortality rate. Prediction models were developed in a competing risk framework for the separate prediction of major haemorrhage and arterial thrombosis over a 90-day follow-up period. Mortality rates could then be explored within defined strata of predicted risk.

9.1 Introduction

The risk of early mortality after ischaemic stroke is largely determined by the severity of the initial stroke in addition to the general health of the patient prior to the stroke. It has been shown that in the case of acute coronary syndrome thrombotic and haemorrhagic complications during recovery have a significant impact on patient mortality rate (Pocock et al., 2010). It is likely that intervening events which occur during recovery from a stroke (e.g., major bleeds and/or thromboses) will also impact the rate of mortality. The influence of acute medical complications and pneumonia in stroke has been investigated though there has been little focus on the separate impact of thrombotic and haemorrhagic events (Grube et al., 2013, Hong et al., 2008). Previous analyses have treated such events as fixed and therefore ignored the natural evolution of these events with time. If it were possible to identify those most likely to suffer a serious complication during recovery then this could be used to help manage the treatment or counselling of stroke patients.

The aims of this chapter are: (i) to explore the impact that early complications have on the risk of mortality; (ii) to determine whether the risk of early complications can

be predicted accurately; and (iii) to illustrate a stratified framework for understanding patient mortality according to grouped predicted risk strata.

9.2 The Virtual International Stroke Trial Archive

The Virtual International Stroke Trial Archive (VISTA) collaborative is run by a group of clinical stroke specialists who are experienced in undertaking clinical trials in stroke. The group has collected and collated datasets of *completed* stroke trials at the patient-level for the purpose of secondary investigation and analyses. It is envisaged that methodological lessons may be learnt through such analyses which could consequently improve the conduct and analysis of future trials (Ali et al., 2007, VISTA Collaborative, 2012). The VISTA-Acute collaboration was established in 2001 offering researchers' access to the patient level data from its repository of stroke trial data. This included a record of patients: medical history; demographic characteristics; laboratory measurements; treatment history; and any endpoints (i.e., primary, secondary and adverse events) recorded during the trial. At the time of writing this thesis, VISTA-Acute was one of six distinct subsections. It is part of strict VISTA policy that no reanalysis of the original randomised trial treatment be allowed. The obtained data extract is therefore stripped of any identifiable treatment codes. There is also no information made available regarding the original trial design. This makes it impossible to identify plausible sources of methodological or clinical heterogeneity. It is therefore important to recognise that any clinical findings made using VISTA data will normally require validation in new or pre-existing datasets which have been reported in detail.

The data are provided by VISTA in a stacked format. Where a unique trial identifier has been supplied it is expected that the investigators will explore heterogeneity. In Chapter 6 a specific form of meta-analysis (type *A*) was introduced: that is an Individual Patient Data Meta-Analysis (IPD-MA) were the original patient level data are available for secondary analysis (Ahmed et al., 2014, Debray et al., 2013). The obtained VISTA extract has the same structure as any other IPD-MA only after the necessary data collection stage. This is an important difference to highlight though and it must be understood exactly what the researcher is blinded to when analysing

VISTA data: The typical format of an IPD-MA will see the investigators search the literature for all studies which meet their inclusion criteria. They will then contact all of the original investigators of these studies inviting them to contribute their raw patient level data. It is implicit that during this process the: patient population; the study design; the original mode of data collection and so on will be known in some detail for each of the individually included studies. Without this information it is not possible to ascribe any *a priori* notions about how sensible it is to pool the data. This is the case with the VISTA *IPD-MA interpretation*: without being able to identify the likely sources of clinical and methodological heterogeneity, any detected statistical heterogeneity that exceeds chance cannot be explained. It is very likely that the individual studies included in the VISTA extract will indeed differ from one another. This is the main drawback of using these data for testing clinical hypotheses, which, for the most part, will be ignored in the analysis that follows.

The VISTA extract made available for analysis in this thesis was updated at various dates. The data were initially obtained on the 8th of October 2009; there were two additional extractions made on the 26th of March 2014 and the 16th of June 2014.

9.2.1 Outcome definitions

The initial extract included a detailed record of adverse events recorded during follow-up with a unique patient ID matching those with adverse events to their baseline data and associated trial endpoints (i.e., mortality and functional outcome). A clinician (Dr. William Whiteley) categorised each of these events into the following broad levels which comprised any: ischaemic strokes (IS); myocardial infarctions (MI); pulmonary embolisms (PE); deep-venous thromboses (DVT); unstable anginas (UA); transient ischaemic attacks (TIA); other thrombotic events; major intracranial haemorrhages (ICH); gastrointestinal haemorrhages (GI); other extra-cranial haemorrhages (ECH); or minor haemorrhages (MH).

More data were requested from VISTA (26/03/2014) who provided an extract with far more detail denoting *any adverse events* as recorded by each trial. This comprised a total of 33,982 records. First this was restricted to those entries that had been coded as “serious” using a unique binary identifier which left 5,801 records. Second, those

adverse event levels categorised as one of the events identified above were mapped onto the new extract, leaving 4,810 rows to search. Of these entries there were 1,527 unique levels. Finally, a text search was carried out so as to limit the length of time spent searching by hand for those *important* events that remained. A series of word fragments were screened against the adverse event list. The word fragments comprised of a number of *key* terms most likely to capture those serious adverse events. By fixing all event data text as upper-case the search undertaken was not case sensitive. The word fragment comprised the following key terms: *haem*; *hemo*; *hem*; *bleed*; *blood*; *ICH*; *infarct*; *cereb*; *throm*; *isch*; *stroke*; *myocard*; *TIA*; *venous*; *recur*; *deep*; *PE*; *embolus*; and *angina*. This reduced the set of events levels to be considered to a total of 383.

For the purpose of this chapter two composite binary events were defined (i) a major haemorrhagic event (i.e., ICH, GI and ECH) and (ii) an arterial thrombotic event (i.e., IS, MI, UA and other thrombotic events). The number of recurrent events was not sufficient to warrant the modelling of multiple thrombotic and/or haemorrhagic events. Following categorisation, all analyses of intermediate adverse events (defined as above) were restricted to the first record only.

Any records of an adverse event denoting a progression of the initial stroke were excluded.

9.2.1.1 Censoring scheme for outcomes

Time to event information was available through a record of the number of days from study start to the observed event. The patients in the VISTA extract had a minimum follow-up of 90 days. The primary endpoints were functional outcome measured on the modified Rankin Scale (mRS) at 90 days. It was assumed that those with as observed mRS had been followed for at least 90 days giving a strict definition of the end of follow-up. Any events beyond this point were right censored.

Of those who suffered an adverse event 28 occurred on the day the study began (i.e., day zero), similarly 30 deaths occurred. To accommodate for these events the observed time was adjusted so that those with an adverse event at time zero had a

random value added to their observed time chosen from a *Uniform* (0, 0.5) distribution. Similarly, for those dead on day zero a random value was chosen from a *Uniform* (0.5, 1) and added to the observed death time. This meant that a patient suffering an adverse event on the day that the study began was randomly chosen a time over the first 12 hours of that day. Likewise, for those patients who died on the day the study began a time across the latter 12 hours of the day was randomly chosen. By setting the parameters of the two uniform distributions in this way the possibility that a patient who suffered an adverse event on the day the study began could also suffer mortality later that same day was allowed.

9.2.2 Baseline characteristics

A record of all the baseline characteristics made available within the obtained VISTA extract is provided in Table 9-1.

Smoking status was recorded in detail with the following unique levels: *ex-smoker*; *habitual smoker*; *occasional smoker*; *answered yes*; *never* and *no*. For the purpose of this chapter levels associated with being a current or previous smoker were collapsed and a binary indicator denoting any smoking versus no smoking was defined. In the case of patient record of ECG scan *abnormal* and *borderline* were combined creating a binary variable.

Detailed records of non-randomised medications were retrieved following a second update of the original extract (16/06/2014). This included the Anatomical Therapeutic Chemical (ATC) Classification System which had been hand coded by a VISTA student directly from each entry recorded in the original clinical report forms per patient. These have been checked and verified by various clinicians and whilst supported by the VISTA group cannot be guaranteed. The ATC codes were used in this chapter to identify previous and current: antiplatelets (including: aspirin [B01AC06]; clopidogrel [B01AC04]; dipyridamole [B01AC07]; and ticlopidine [B01AC05]) and anticoagulants (including: heparin [B01AB01]; warfarin [B01AA03]; and dicoumarol [B01AA01]). Each medication was coded as a binary yes/no variable. A record of the start time and stop time measured in days from the beginning of the trial was available. To define the variables: *prior antiplatelet* and

prior anticoagulant, any of the treatment records characterised by a positive entry on the prior medication variable or else those with a negative start time (used to denote pre-origin) were selected using the aforementioned ATC codes for various antiplatelets and anticoagulants. Current treatments were recorded throughout trial follow-up and are therefore strictly speaking "time-varying covariates". To avoid introducing any further complexity to the prediction models though, these were treated as fixed and restricted to only those administered within the first week (i.e., 7 days) of the study origin.

Table 9-1 Baseline characteristics and outcome event information for VISTA data extract.

Variable	Description
Patient characteristics and medical history	
Trial	Anonymous trial ID (categorical variable, 1 to 7)
Age	Measured in years at baseline
Male	Male=1/female=0
Weight	Measured at baseline in kgs
Systolic BP	Measured at baseline in mmHg
Diastolic BP	Measured at baseline in mmHg
Smoker*	Detailed record, coded as current or previous=1/none=0
History of hypertension	Present=1/Absent=0
History of previous stroke	Present=1/Absent=0
History of TIA	Present=1/Absent=0
History of diabetes mellitus	Present=1/Absent=0
History of MI	Present=1/Absent=0
History of CHF	Present=1/Absent=0
History of atrial fibrillation	Present=1/Absent=0
Detail of stroke on admission	
NIHSS	Measured at baseline
Type or stroke	Ischaemic=2/haemorrhagic=1/unknown=0
Hemisphere	Stroke in right hemisphere=1/ left=0
ECG results	Recorded as Abnormal=2/Borderline=1/Normal=0
Cause of initial stroke	Recorded as large vessel/cardioembolic/lacunar/other
Medication	
Prior antiplatelets*	Present=1/Absent=0
Prior anticoagulants*	Present=1/Absent=0
Current antiplatelets*	Present=1/Absent=0
Current anticoagulants*	Present=1/Absent=0
rtPA administered	Present=1/Absent=0
Outcome	
Major haemorrhage*	Intracranial, gastrointestinal or other extra-cranial (1/0)
Time of above	Recorded in days
Minor haemorrhage*	Any minor bleeds (1/0)
Time of above	Recorded in days
Arterial thrombosis*	Ischaemic stroke, myocardial infarction, unstable angina, or other thrombosis (1/0)
Time of above	Recorded in days
Venous thrombosis*	Pulmonary embolism or deep venous thrombosis (1/0)
Time of above	Recorded in days
mRS	Recorded at 90 days (ordinal measure ranging from 0 to 6)
Mortality	Present=1/Absent=0
Time of above	Recorded in days

Note: (*) defined for this thesis. Abbreviations: CHF – Coronary Heart Failure; TIA – Transient Ischaemic Attack; NIHSS – National Institute of Health Stroke Score; mRS – modified Rankin Score; rtPA – recombinant tissue plasminogen activator; ECG – electrocardiogram; BP – blood pressure; MI – Myocardial Infarction

9.3 Methods: Time-varying Cox PH Model

The impact made on the rate of mortality by an adverse event that occurs during follow-up can be measured by introducing the variable as a time-varying covariate within a Cox Proportional Hazards (PH) modelling framework. This is a simple extension of the *fixed*-covariate Cox PH model which was defined in Chapter 7. The terms "fixed" and "varying" are introduced here so as to distinguish the two covariate types, with *fixed*-covariate Cox PH models denoting the standard Cox model. The time-*varying* Cox PH model accommodates for the inclusion of covariates which can change value over time. This is achieved by allowing multiple entries per patient conditional on the time-varying covariates change (Hosmer Jr et al., 2011). All time-varying covariates considered in this chapter are categorical: each indicating the presence or absence of an associated *complication* i.e., an arterial thrombosis and/or a major haemorrhage. A general vector of patient characteristics, $\boldsymbol{z}(t)$, is specified such that each of the included elements can be defined as a function of time which includes *fixed*-covariates as a special case,

$$h(t \mid \boldsymbol{z}(t)) = h_0(t) \exp\left\{\boldsymbol{\beta}^T \boldsymbol{z}(t)\right\}. \qquad \textbf{(9.1)}$$

Parameter estimates are obtained in exactly the same way as the *fixed*-covariate Cox PH model by maximising the partial log-likelihood. Coding a time-varying covariate for a Cox PH model requires some caution as a degree of manipulation is required so as to format the data correctly. Imagine using a stopwatch to record the track times for a series of runners competing in the 400m sprint. Suppose that if a runner stumbles during the course of their run, a *split* will be used thus marking their stumble-time in addition to their track-time. This is precisely what is required in the case of a time-varying covariate. If an adverse event occurs during follow-up: the time is split at the *stumble-time* creating two entries for that patient; otherwise only one entry is required. The interpretation of the effect estimates differ from the *fixed*-covariate fit (Altman and De Stavola, 1994). Under the *fixed*-covariate Cox PH model each effect estimate is made with respect to a covariate measured at time zero. For the time-dependent Cox PH model the effect estimates are obtained on entry or at any other time during follow-up, under the PH assumption.

9.4 Methods: Modelling Competing Events

As was noted in section 9.2.1, patient recovery is complex and subject to many additional adverse events. Composite events were created in the VISTA data so as to simplify the problem as well as increase the power through the number of events per variable (Steyerberg, 2009). Restricting to the two adverse *composite events* and mortality observed over follow-up, a standard competing event set-up is described. Patient recovery is assumed to follow a particular pathway as seen in Figure 9-1. All patients enter the study alive (the *initial* state) then, over the course of follow-up, one of four events are observed: (1) arterial thrombosis; (2) major haemorrhage; (3) death; or (4) right censoring (i.e., alive at the end of follow-up).

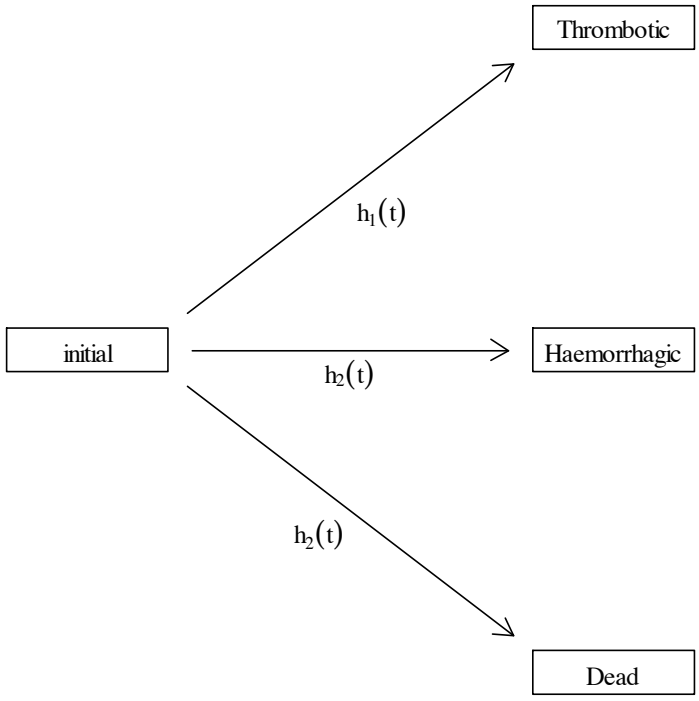


Figure 9-1 The competing event set up with endpoints: arterial thrombosis; major haemorrhage; and death.

It is noted that more complex multi-state processes can be envisaged with patients passing through states (e.g., initial → arterial thrombosis → major haemorrhage) possibly with an absorbing state, e.g., death. In the context of this chapter only the

competing event framework described in Figure 9-1 will be considered. All analysis is restricted to first event data only.

Outcomes of this form require an appropriate method of analysis. One such approach works by extending the standard methods used in survival analysis to incorporate such complex outcome structures. In general, patient outcome information is fully described by a record of *time* to event, T, and a record of the *type* of event, D, which takes on any real valued number between 0 and k (Note that 0 typically denotes a *censored* event). Therefore the observed time, T, is the minimum time till one of the events $0, 1, 2, \ldots, k$ occur. The specific set up described in Figure 9-1 implies that k takes on the real values 1, 2 and 3 representing: arterial thrombosis; major haemorrhage; and death respectively.

9.4.1 A competing event

Concepts unique to survival analysis were introduced in Chapter 8: Patients are followed for a fixed period of time during which any events of interest are recorded (e.g., recurrent stroke or death). The idea of censoring was discussed as not every patient will have an observed time-to-event over follow-up; this means that a proportion will either be censored during follow-up or else censored at the end of follow-up (i.e., right censored). They will however have censoring times. This introduces one of the essential assumptions in survival analysis: that the event distribution and the censoring distribution are independent of one another. In the case of competing events though an additional reason for censoring is introduced. The following analogy is offered as a way to illustrate this.

A game of outdoor tennis is to be played between two opponents. Match day comes but immediately before the two adversaries take to the court there is a freak thunder storm calling halt to the contest before it can even begin. The match score was the event of interest however this was precluded by the occurrence of a separate and exclusive *competing event*: the thunder storm. When studying the elderly or frail who are subject to a far greater number of multiple morbidities there is an abundance of competing events which can occur. Despite this the analyses traditionally adopted by investigators will often ignore or inappropriately handle these issues, typically

censoring on the competing event and thus obtaining incorrect predicted risks (Koller et al., 2012, Wolbers et al., 2009). The key issue is that by assumption when censoring a patient it is argued that the patient could go on to develop the event of interest in the future. However, when the reason for censoring precludes the event of interest then this assumption no-longer holds and the error of making it will introduce a bias into prediction (Putter et al., 2007).

Throughout this chapter the events of primary interest are *major haemorrhages* and *arterial thromboses*; with death playing the role of the freak thunderstorm calling time on the occurrence of either of these adverse events.

9.4.2 Analysing competing risk data

9.4.2.1 Cause specific hazards

The hazard function is fundamental in the analysis of survival data; however, the hazard function as defined in Chapter 8 does not incorporate competing events. An adaptation of this function is defined which describes the hazard of suffering the event of interest (e.g., arterial thrombosis) in the presence of some competing event. This is the *cause-specific hazard function* (Putter et al., 2007):

$$h_k(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t, D = k \mid T \ge t)}{\Delta t} \quad \textbf{(9.2)}$$

Here D denotes the possibly censored outcome, e.g., arterial thrombosis ($k = 1$, with $h_1(t)$) and death ($k = 2$, with $h_2(t)$). It follows that by integrating over time the cumulative cause specific hazard is obtained and similarly the marginal survival function can be obtained as $\exp(-H_k(t))$.

When working with competing events the cumulative incidence function provides a non-parametric estimate of the probability of failure from any cause:

$$I_k(t) = \int_0^t \lambda_k(s) S(s) ds \quad \textbf{(9.3)}$$

In the absence of censoring, the cumulative incidence function for the k^{th} event can be regarded as the evolution of the proportion of the k^{th} event over time as observed

in the sample (Wolbers et al., 2014). Adjustments must be made when censoring has occurred (Putter et al., 2007). An estimate for equation (9.3) is given by:

$$\hat{I}_k(t) = \sum_{j:t_j \le t} \hat{h}_k(t_j)\hat{S}(t_{j-1}) \tag{9.4}$$

The cumulative incidence function describes the joint probability of failing from the k^{th} cause having survived up until some given time point, t_{j-1} as estimated by the Kaplan-Meier estimator (van Houwelingen and Putter, 2012). An estimate for the cause specific hazard is given by,

$$\hat{h}_k(t_j) = {}^{d_{kj}}\!/_{n_j}, \tag{9.5}$$

which, for discrete time, is simply the ratio of the number of observed events at time t over the number that remain within the risk set.

9.4.2.2 Regression models for competing risk

The two most commonly adopted methods for adjusting for multiple covariates in a competing risk analysis are: (i) cause-specific Cox PH models and (ii) the cumulative incidence Fine and Gray (FG) modelling approach. Each method is characterised by the way the event causes are handled and the necessary definition of the risk set, i.e., those that remain at risk from the event of interest (Lau et al., 2009).

With two exclusive endpoints, i.e., an event of interest and the competing event, two Cox PH models are fit modelling the cause specific hazards for: (i) the event of interest, which censors on the competing event as well as any other right censored observations; and similarly (ii) a model for the competing event itself, this time censoring on the event of interest and any right censored observations:

$$h_k(t \mid \boldsymbol{z}) = h_{k,0}(t)\exp\left(\boldsymbol{\beta}_k^T \boldsymbol{z}\right). \tag{9.6}$$

The cause-specific hazard approach therefore involves modelling the specific hazard from each failure using standard Cox PH models. Subjects who suffer any observed events (i.e., the competing event or the event of interest) are removed from the risk set used to estimate the respective hazard (Lau et al., 2009). In order to produce absolute risk estimates the following expression must be used which combines each of the cause-specific hazard models obtained for each mode of failure,

$$I_1(t) = \int_0^t h_1(s)\exp\left(-\int_0^s (h_1(u) + h_2(u))du\right)ds\ . \qquad \textbf{(9.7)}$$

This avoids introducing the bias associated with a standard application of the Cox PH model (Wolbers et al., 2009).

Fine and Gray suggested a different approach which involves direct regression on the cumulative incidence via the subdistribution hazard (Fine and Gray, 1999). The subdistribution hazard is based on the complement to the incidence function for the k^{th} event, i.e., the absolute estimate of a subject surviving without suffering the event.

This is analogous to the survival function which can be transformed as the hazard, or in the case of competing risks, the subdistribution hazard,

$$\overline{h}_k(t) = -\frac{d\log\left(1 - I_k(t)\right)}{dt}, \qquad \textbf{(9.8)}$$

with $I_k(t)$ as defined in equation (9.3). Putter *et al*. note that the subdistribution hazard is uniquely different from the cause-specific hazard in how the risk set is defined (Putter et al., 2007). In the case of the cause-specific hazards, failure from any event will result in a removal from the corresponding risk set. For the subdistribution hazard though, any failures from the *competing event* are retained within the risk set (Lau et al., 2009).

Fine and Gray suggested direct regression on the subdistribution hazard through a model which assumes proportional subdistribution hazards,

$$\bar{h}_k(t \mid \boldsymbol{z}) = \bar{h}_{k,0}(t)\exp\left(\boldsymbol{\beta}_k^T \boldsymbol{z}\right). \quad \textbf{(9.9)}$$

All aspects of model development outlined in Chapter 2 apply in the development of a prediction model in the presence of competing events. The Fine and Gray approach has been highlighted as appealing from the perspective of prediction, whilst modelling the cause-specific hazards is deemed more suitable for etiological research (Wolbers et al., 2009, Lau et al., 2009).

9.4.2.3 Model development – assessing non-linearity

Restricted cubic splines (RCS) have been used throughout this thesis to explore the plausible existence of non-linear relationships between continuous predictors and outcomes. This process is complicated when multiple imputation methods have been used to account for missing data. Typically the placement of the knots will be at specific values which correspond to the percentiles of the observed sample distribution, for example with four knots, the 5th, 35th, 65th and 95th percentiles would be used (Harrell, 2001). Alternatively, their exact placement can be specified, though this will usually require some clinical justification. Nevertheless, imputing missing values multiple times generates several complete and conditionally plausible datasets, each of which will have slightly different covariates distributions and thus slightly different knot positions specified by the percentiles of the distribution. A more explicit and practical approach was demonstrated by Vergouwe *et al*., with the adoption of a *majority multiple fractional polynomial* selection process (Vergouwe et al., 2010) whereby the most frequently selected transformation out of the *m* imputed datasets is chosen for the final pooled model fit.

Fractional polynomials (FP) offer an alternative method for the modelling of non-linear predictor outcome associations (Royston and Altman, 1994). Here a reduced set of powers are specified, e.g., $P = [-2, -1, -0.5, 0, 0.5, 1, 2, 3]$ where 0 denotes the natural logarithm, *ln* (*X*), which necessarily restricts the predictor to be positive and therefore enforces the required scaling, $x^* = x + \delta$, for the random variable x with some constant δ such that x^* is positive (Sauerbrei and Royston, 1999). For a first degree fractional polynomial a single transformation is chosen from this small set of

exponents outlined in P. Selection involves iterating though a table of deviance statistics which at each stage contrast the proposed transformation (i.e., $p \neq 1$) with the simple linear fit (i.e., $p = 1$). More flexible fits can be allowed by adding a larger number of transformations of x from the same set of exponents, P. Note that if a first-degree fractional polynomial is of the form $\beta_0 + \beta_1 x^{p1}$, then a two-degree fractional polynomial is of the form $\beta_0 + \beta_1 x^{p1} + \beta_2 x^{p2}$. In practice, a two-degree fractional polynomial is adequate to capture most transformations (Royston et al., 1999). The general expression for the m^{th} degree FP is given by,

$$g\{E(Y)\} = \beta_0 + \sum_{j=1}^{m} \beta_j f_j(x), \qquad \textbf{(9.10)}$$

where,

$$f_j(x) = \begin{cases} x^{p_j} & \text{if } p_j \neq p_{j-1} \\ f_{j-1}(x)\ln(x) & \text{if } p_j = p_{j-1} \end{cases}.$$

Routines for selecting a fractional polynomial transformation for continuous predictors in an FG model are not currently available. However, in practice the two approaches (i.e., cause-specific Cox PH models and FG models) give similar results (Wolbers et al., 2009) therefore in this chapter non-linear transformations were selected under separate cause-specific hazard models for arterial thrombosis and major haemorrhage before being applied within the Fine and Gray subdistribution hazard framework.

9.4.2.4 Model development – variable selection

The variables to include in each of the multivariable clinical prediction models are restricted by the availability of patient variables provided in the original VISTA extract. Issues with overfitting were discussed in Chapter 2. However, with few variables to choose from no further selection was deemed necessary. In each case a full prediction model was fit with no screening of significance.

9.4.2.5 Model performance – competing event models

For prediction models, discrimination summarises how well a given model or score separates observed events in follow-up based on a weighted sum of characteristics measured at baseline. A standard statistic in this context is the *c*-index: the probability of correctly ordering a given pair of patients, one with an event and one without. The patient with the event should be assigned a greater score or predicted risk than the patient without. With competing events it is necessary to make an alteration to the definition of *possible pairs*. One such definition is that which was proposed by Wolbers *et al*. who suggest that a relevant pairing can be made between those who suffer the event of interest at some given time point and those patient still at risk (Wolbers et al., 2009). This requires that any competing events are retained within the risk set whilst those who suffer either the event of interest or else those censored (whichever occurs first) are removed at the corresponding time during follow-up. This definition closely corresponds to the risk set implicitly used during Fine and Gray regression on the subdistribution hazard.

Absolute estimates of patients' predicted risk can be obtained using a Fine and Gray model. Typically, a non-parametric estimate of the observed absolute risk for a given grouping of predicted risk made through a Cox PH model (e.g., predicted categorised into equally sized deciles) is obtained using the Kaplan-Meier estimator (van Houwelingen, 2000). With competing events the cumulative-incidence function is used in place of the Kaplan-Meier estimator (Wolbers et al., 2009).

9.4.3 Between trial heterogeneity

The trial level data made available for analysis within the obtained VISTA extract comprised data from seven individual trials. The predictor-outcome associations could differ between trials; therefore between trial heterogeneity was explored. Formal tests for heterogeneity are presented in an attempt to assess departures from a single common effect (Debray et al., 2013).

9.4.4 Missing data

A multiple imputation method is used to account for any observed missingness in the record of patients' baseline characteristics in the VISTA extract. This should reduce

the impact of the biases associated with complete-case analyses should the data be missing at random (Vergouwe et al., 2010).

The structure of the dataset plays an important role in how missingness should be explored. It is implicit that if between-trial heterogeneity is of interest then *Trial* itself must be incorporated within the imputation process. Resche-Rigon *et al*. discuss an across studies approach for dealing with systematically missing values in the context of IPD-MAs (Resche-Rigon et al., 2013). Often entire covariates will be missing from certain studies, however; the information between trials may be utilised. Treating *Trial* as random in this multi-level approach would be one way of doing a one-step imputation with stacked datasets. The approach taken in this chapter though is to regard the process as one which might be encountered in practice with *Trial* being treated as a stratification factor and separately implementing multiple imputation within each of the identifiable datasets. First prediction models will be developed in a single trial where missing data is handled using multiple imputation. Second the developed models will be evaluated within the remaining *Trial* cohorts, where again for each of the identifiable datasets, missing data will be handled, generating multiple imputed datasets.

It is recommended that when imputing missing data for survival analysis instead of using the raw time variable during imputation, a cumulative hazard estimate (e.g., the Nelson–Aalen estimate) should be used in its place (White and Royston, 2009). The cause specific cumulative hazards for death, arterial thrombosis and major haemorrhage were therefore used during imputation.

9.4.5 Sensitivity Analyses

Concomitant treatment with rtPA was reported in two out of seven trials. It is not clear whether this is a misreporting or true absence of treatment with rtPA. This was explored through a sensitivity analysis.

9.5 Results

Of the 10574 patients made available for analysis 10003 had an ischaemic or unknown stroke on entry. There was evidence to suggest that 190 of these patients experienced a progressive stroke and where therefore removed – leaving a total of 9834 patients for analysis. Over the course of a 90-day follow-up period 465 (5%) and 404 (4%) of these patients developed an arterial thrombosis and a major haemorrhage respectively.

9.5.1 Baseline characteristics and outcomes

Patient summary characteristics split by trial are provided below (see Table 9-2 and Table 9-3). The median patient age varied between trials, ranging from as young as 67 to as old as 74. Similarly, median stroke severity, as measured using the NIHS Score, ranged from as low as 11 to as high as 17. From these tables it is evident that patient characteristics differed from trial to trial according to both the prevalence of various risk factors measured at baseline in addition to the incidence of outcomes recorded in follow-up.

A non-parametric estimate of the hazard functions for arterial thrombosis, major haemorrhage and death is provided in Figure 9-2 using the `muhaz` library in R (Kenneth and Gentleman, 2010). There is a notable positive skew in each distribution suggesting that patients are at greatest risk immediately following their initial stroke. The hazard rate for major haemorrhage is larger than that of arterial thrombosis but declines rapidly over the first week whilst the hazard rate of arterial thrombosis has a steadier decline over the first 20-days or so after which point the two rates are about the same.

9.5.2 Missing data

The proportion and frequency of observed missingness for each variable is summarised in Table 9-2 and Table 9-3. Of the seven trials, *Trial 1* was both the largest and the most complete. For a number of trials some variables were missing entirely – either because they were never recorded or else never provided to the VISTA collaborative. For *Trials 2, 4, 5, 6* and *7* there was no record of patients receiving rtPA and no record of patients with any prior Coronary Heart Failure

(CHF). There was no record of: prior Transient Ischaemic Attack (TIA) for *Trials 4* and *5*; of prior stroke for *Trials 6* and *7*; and finally of electrocardiogram (ECG) or of any arterial thrombotic outcomes in follow-up for *Trial 3*.

Joint missingness was explored for each of the seven obtained trials using cluster plots (see section 9.7 Appendix A on page 332). In what follows those completely missing variables are excluded from interpretation: Of those recruited to *Trial 1*, 4640 (94%) had completely observed baseline data, 286 had one missing value (6%) and 17 (<1%) had two or more missing values. For *Trials 2* to *6* the median proportion of completely observed baseline data was 86% (with Inter Quartile Range [IQR] 81% to 89%). More detail regarding the frequency of NAs per observation is provided in Appendix A.

A multiple imputation model was implemented within each of the trial datasets, conditioning in each case on the dependency specific to the corresponding trial. This was implemented in R using the `mice` library (van Buuren and Groothuis-Oudshoorn, 2011). The adopted model included: all of the variables listed in Table 9-2 and Table 9-3; indicator variables for the events of interest in follow-up; and the estimated hazard function for 90 day arterial thrombosis, major haemorrhage and mortality. A new variable was defined so as to account for those instances of missing variables for vascular events. Therefore, any record of: previous stroke, TIA; and MI were merged into one newly defined composite variable denoting *any vascular event*. Predictive mean matching was used for continuous variables, binary logistic regression for binary categorical variables and proportional odds models for ordinal categorical variables. Missing data were imputed multiple times generating ten complete datasets per trial.

For the most part, *Trial 3* is excluded from the analysis conducted in this chapter due to the complete lack of any arterial thrombotic adverse events in follow-up. Explicit reference will be made to its use in the case of haemorrhagic events.

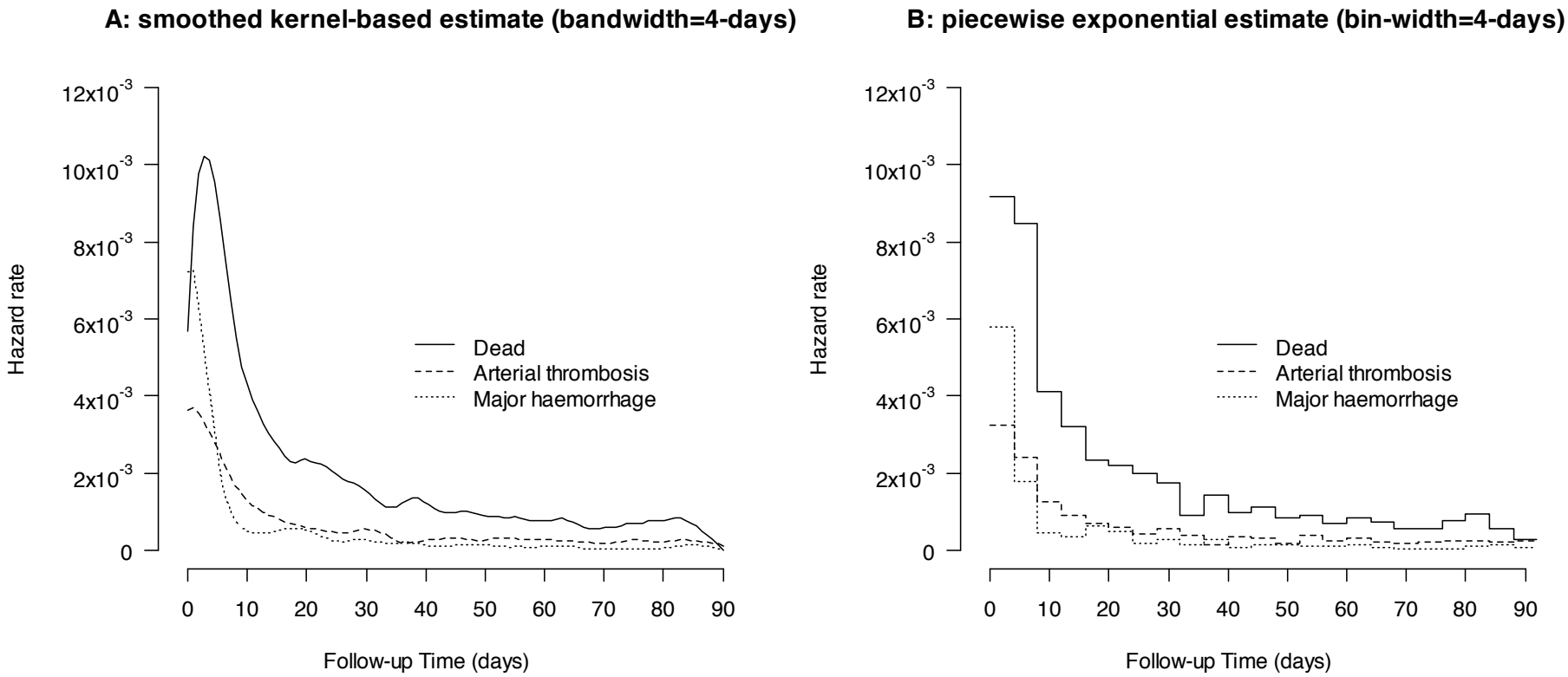


Figure 9-2 Estimates of the instantaneous rate of failure: (A) smoothed kernel estimate with bandwidth set at 4-days; and (B) piecewise exponential estimate with bin-width set at 4-days.

Table 9-2 Patient characteristics for trials 1 through to 4. (Note: acronyms are as denoted in Table 9-1). For continuous measurements the median and inter-quartile range is quoted; for categorical measurements the frequency and % is provided (continued in Table 9-3).

	Trial 1 (N = 4943)		Trial 2 (N = 574)		Trial 3 (N = 825)		Trial 4 (N = 1289)	
Characteristic	**Measure**	**Missing**	**Measure**	**Missing**	**Measure**	**Missing**	**Measure**	**Missing**
Age, years	71 (60-78)	-	74 (66-81)	-	73 (63-79)	1 (<1)	72 (62-78)	-
Gender, male	2713 (55)	-	274 (48)	-	428 (52)	1 (<1)	683 (53)	-
Weight, kg	75 (65-85)	2 (<1)	73 (62-84)	1 (<1)	75 (66-85)	79 (10)	75 (65-86)	10 (1)
SBP (mmHg)	154 (138-172)	14 (<1)	151 (135-172)	16 (3)	155 (140-172)	54 (7)	152 (135-170)	1 (<1)
DBP (mmHg)	82 (72-94)	15 (<1)	80 (69-91)	16 (3)	80 (72-90)	55 (7)	80 (70-90)	2 (<1)
Smoker, yes	2167 (44)	10 (<1)	303 (53)	-	769 (93)	47 (6)	322 (25)	119 (9)
Atrial Fibrillation	1270 (26)	-	152 (26)	-	216 (26)	47 (6)	319 (25)	119 (9)
Hypertension	3663 (74)	-	428 (75)	-	548 (66)	47 (6)	873 (68)	119 (9)
Previous stroke	978 (20)	-	125 (22)	-	83 (10)	47 (6)	624 (48)	75 (6)
TIA	409 (8)	279 (6)	109 (19)	-	50 (6)	47 (6)	0 (0)	1289 (100)
Diabetes	1135 (23)	-	139 (24)	-	180 (22)	47 (6)	315 (24)	119 (9)
Myocardial Infarction	641 (13)	-	87 (15)	-	103 (12)	47 (6)	288 (22)	119 (9)
CHF	467 (9)	-	0 (0)	574 (100)	38 (5)	47 (6)	0 (0)	1289 (100)
NIHSS	12 (8-17)	-	17 (13-20)	-	15 (11-18)	48 (6)	11 (7-17)	-
Hemisphere	2625 (53)	1 (<1)	264 (46)	5 (1)	376 (46)	50 (6)	655 (51)	1 (<1)
ECG	3206 (65)	-	395 (69)	91 (16)	0 (0)	825 (100)	703 (55)	19 (1)
Prior antiplatelet	1606 (32)	-	92 (16)	-	305 (37)	-	535 (42)	-
Current antiplatelet	3163 (64)	-	146 (25)	-	509 (62)	-	0 (0)	-
Prior anticoagulant	332 (7)	-	49 (9)	-	70 (8)	-	207 (16)	-
Current anticoagulant	1307 (26)	-	183 (32)	-	279 (34)	-	0 (0)	-
rtPA	1915 (39)	-	0 (0)	-	483 (59)	47 (6)	316 (25)	-
Outcomes by 90 days								
Arterial thrombosis	311 (6)	-	16 (3)	-	0 (0)	-	53 (4)	-
Major haemorrhage	236 (5)	-	23 (4)	-	12 (1)	-	49 (4)	-
Dead	581 (12)	-	101 (18)	-	129 (16)	-	142 (11)	-

Table 9-3 Patient characteristics for trials 5 through to 7 as well as an overall trials summary. (Note: acronyms are as denoted in Table 9-1). For continuous measurements the median and inter-quartile range is quoted; for categorical measurements the frequency and % is provided.

	Trial 5 (N = 1419)		Trial 6 (N = 383)		Trial 7 (N = 401)		Total (N = 9834)	
Characteristic	**Measure**	**Missing**	**Measure**	**Missing**	**Measure**	**Missing**	**Measure**	**Missing**
Age, years	72 (63-79)	1 (<1)	67 (59-73)	-	68 (61-75)	-	71 (61-78)	2 (<1)
Gender, male	803 (57)	1 (<1)	245 (64)	-	227 (57)	-	5373 (55)	2 (<1)
Weight, kg	72 (63-80)	71 (5)	73 (65-82)	-	74 (65-84)	-	75 (65-85)	163 (2)
SBP (mmHg)	160 (140-180)	2 (<1)	150 (140-170)	-	154 (140-169)	-	154 (138-171)	87 (1)
DBP (mmHg)	86 (79-97)	2 (<1)	90 (80-100)	1 (<1)	85 (77-90)	-	82 (72-93)	91 (1)
Smoker, yes	307 (22)	191 (13)	1 (<1)	-	233 (58)	2 (<1)	4102 (42)	369 (4)
Atrial Fibrillation	414 (29)	191 (13)	66 (17)	-	28 (7)	152 (38)	2465 (25)	509 (5)
Hypertension	772 (54)	191 (13)	133 (35)	-	98 (24)	152 (38)	6515 (66)	509 (5)
Previous stroke	628 (44)	133 (9)	0 (0)	383 (100)	0 (0)	401 (100)	2438 (25)	1039 (11)
TIA	0 (0)	1419 (100)	31 (8)	-	2 (<1)	152 (38)	601 (6)	3186 (32)
Diabetes	259 (18)	191 (13)	63 (16)	-	23 (6)	1 (<1)	2114 (21)	358 (4)
Myocardial Infarction	200 (14)	191 (13)	18 (5)	-	3 (1)	152 (38)	1340 (14)	509 (5)
CHF	0 (0)	1419 (100)	0 (0)	383 (100)	0 (0)	401 (100)	505 (5)	4113 (42)
NIHSS	12 (8-18)	4 (<1)	13 (7-18)	15 (4)	12 (8-16)	-	12 (8-18)	67 (1)
Hemisphere	711 (50)	7 (<1)	185 (48)	15 (4)	193 (48)	23 (6)	5009 (51)	102 (1)
ECG	814 (57)	37 (3)	215 (56)	9 (2)	146 (36)	3 (1)	5479 (56)	984 (10)
Prior antiplatelet	415 (29)	-	76 (20)	-	31 (8)	-	3060 (31)	-
Current antiplatelet	0 (0)	-	99 (26)	-	196 (49)	-	4113 (42)	-
Prior anticoagulant	42 (3)	-	6 (2)	-	2 (<1)	-	708 (7)	-
Current anticoagulant	0 (0)	-	120 (31)	-	174 (43)	-	2063 (21)	-
rtPA	0 (0)	-	0 (0)	-	0 (0)	-	2714 (28)	2824 (29)
Outcomes by 90 days								
Arterial thrombosis	55 (4)	-	16 (4)	-	14 (3)	-	465 (5)	-
Major haemorrhage	43 (3)	-	29 (8)	-	12 (3)	-	404 (4)	-
Dead	185 (13)	-	40 (10)	-	25 (6)	-	1203 (12)	-

9.5.3 Measuring the impact of adverse events on mortality

The median time to an arterial thrombotic event across all trials (excluding *Trial 3*) was 11 days (IQR: 4 to 32) and 3 days (IQR: 1 to 15) for a major haemorrhagic event. Of those who suffered a major haemorrhage during follow-up 44% were dead by the end of the 90 day follow-up, similarly of those who suffered an arterial thrombotic event, 51% died, whilst 15% of those without a major haemorrhage and 15% of those without an arterial thrombosis died by the end of the 90-day follow-up.

Note that in setting up the time-varying covariates it was noted that on occasion a death and an adverse event were recorded as having occurred on the same day. An approach implemented by Frank Harrell in the `ie.setup` function in his `rms` library is to subtract some random number chosen from a *Uniform* (0, 1) which thus breaks such ties (Harrell, 2013). The approach used here therefore stipulates that death can follow an adverse event but cannot occur exactly at once.

The risk of 90-day all-cause mortality was assessed with an adjusted time varying covariate Cox proportional hazards model. The risk of mortality was predicted for each of the patients using a simple pre-existing prediction model based on patient age and stroke severity (baseline NIHS Score) previously developed and evaluated within the VISTA data to predict the risk of three month mortality amongst stroke patients with acute cerebral ischemia (König et al., 2008, Weimar et al., 2004). The obtained linear predictor (i.e., on the log odds scale) was entered into the time updated Cox PH model along with the two time-varying binary covariates: arterial thrombosis and major haemorrhage. The possibility of a non-linear association between outcome and the log odds of mortality was allowed through a restricted cubic spline with five knots. Finally, the model was stratified by *Trial* therefore incorporating a different baseline hazard for each level of *Trial* whilst assuming common effects for all other covariates across the unique *Trial* levels.

Table 9-4 below summarises the constant relative effect upon the 90 day hazard of mortality and the proportional hazard assumption. The occurrence of any major haemorrhagic events or any arterial thrombotic events during recovery was independently associated with the rate of mortality even after adjusting for patients

underlying predicted risk of mortality. The quoted hazard ratios suggest an increase in the rate of mortality, however when inspecting the Schoenfeld residual plots (Figure 9-3) a more complex association is apparent: an interaction with time. A test for the correlation between rank ordered time and the scaled Schoenfeld residuals gave P-values <0.0001 and 0.0065 for major haemorrhage and arterial thrombosis respectively. The visible departure from the constant effect estimate assumed by the time-independent Cox PH model implies that the effect of an early adverse event has a larger relative effect on mortality than a late adverse event. The impact of a major haemorrhage and arterial thrombosis on the risk of mortality within 90-days of acute ischaemic stroke cannot therefore be summarised by two constant relative hazards.

It could be argued that this dependency on time holds some biological plausibility when it is considered that those who suffer a serious adverse event not long after their initial stroke have had less time to recover and thus the impact on mortality is at its greatest – though this is speculation (Figure 9-3).

Table 9-4 Impact of intervening events on mortality rate, fit as time varying covariates with adjustment for the underlying risk of mortality in a stratified Cox PH model (see text for details).

Intervening event	Number dead	Adjusted HR	95%CI	P-value
Arterial thrombosis				
Absent	1248	1.00	-	
Present	235	12.94	11.18 to 14.99	<0.0001
Major Haemorrhage				
Absent	1312	1.00	-	
Present	171	5.60	4.73 to 6.62	<0.0001

Note –26 patients were deleted due to missingness in the record of patient age and NIHSS

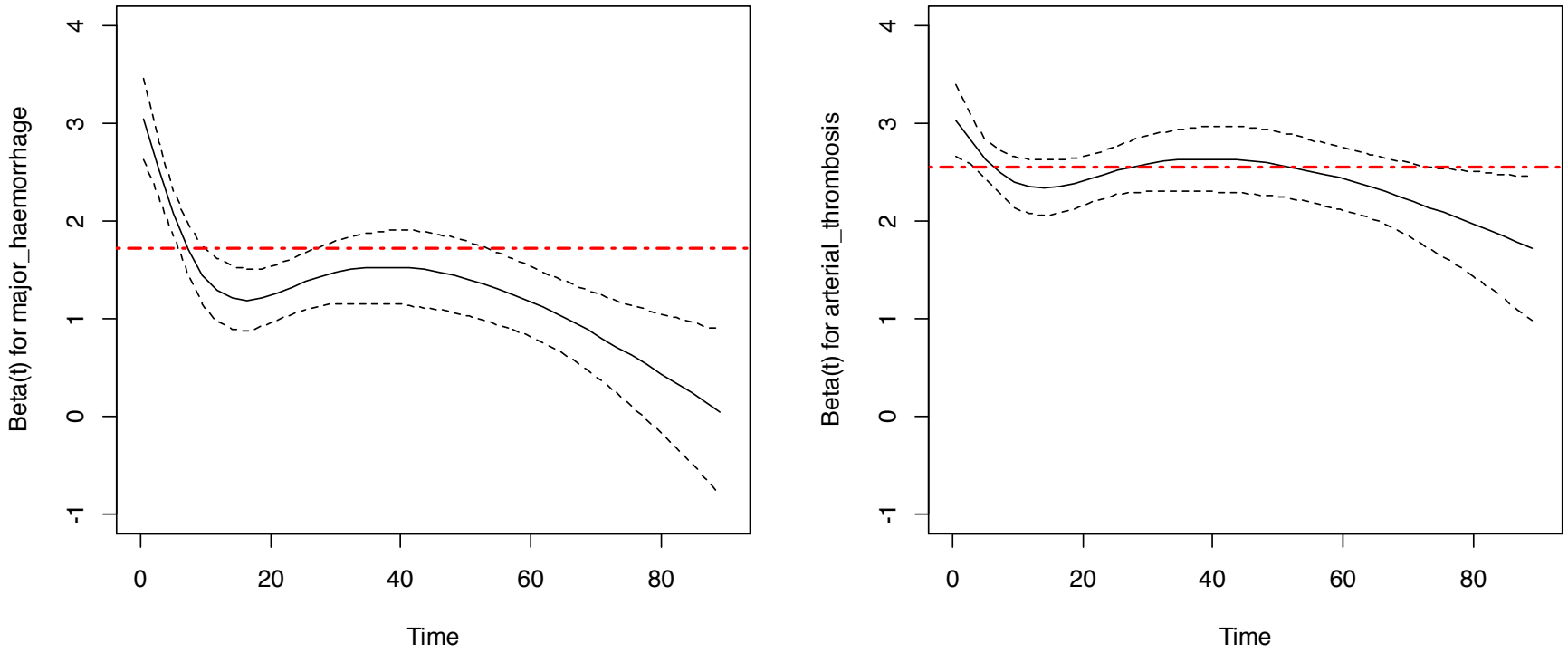


Figure 9-3 Spline fit through Schoenfeld residuals for the time-varying covariates: arterial thrombosis and major haemorrhage. The broken red horizontal line shows the constant estimated effect under the proportional hazards assumption from the Cox PH model on the log hazard scale.

9.5.4 Prediction of early complications

9.5.4.1 Univariable associations for adverse events

Not all of the baseline patient characteristics were observed for each of the trials (Table 9-2 and Table 9-3). An estimated predictor-outcome association cannot be obtained in each of the available trials. Exploring heterogeneity on the univariable level therefore necessarily depends upon the associated number of individual trial estimates made. Recall that no covariate effects on the risk of arterial thrombosis could be obtained for *Trial 3* as there were no records of any such events. Each of the univariable associations were obtained under the Fine and Gray model with death and the opposing adverse event regarded as *competing events* in the case of arterial thrombosis and major haemorrhage respectively per trial. All continuous covariates were fit as simple linear terms. Each estimate was pooled over the 10 imputed datasets using Rubin's rules; these individual trial estimates were then pooled across the available trials using both fixed effect and random effects meta-analysis. A formal assessment of heterogeneity was made using the Q-statistic. Between trial heterogeneity was regarded as statistically significant with a P-value <0.05.

None of the covariate associations with *arterial thrombosis* had a statistically significant level of between trial heterogeneity (Table 9-5); although there was statistically significant between trial heterogeneity present in the estimates of the impact of NIHSS on *major haemorrhage* (Table 9-6).

The pooled estimates for the risk of *arterial thrombosis* were similar whether pooled using a fixed effect or a random effects meta-analytic approach. There was evidence to suggest that those patients who were: older; with worse index strokes (i.e., larger NIHS Scores); with a history of atrial fibrillation; hypertension; any previous vascular events; diabetic; with an abnormal ECG scan; and who were on prior antiplatelets or anticoagulants were at an increased risk from suffering an arterial thrombotic event. Similarly being male and having a smaller weight was associated with a decreased risk from arterial thrombosis.

The pooled estimates for the risk of *major haemorrhage* were similar irrespective of the method of pooling (i.e., fixed or random). There was evidence to suggest that

patients who were: older; with worse index strokes; a history of atrial fibrillation; had an abnormal ECG scan; and who were on prior antiplatelets or anticoagulants were at an increased risk from suffering a major haemorrhage. Similarly, a history of having smoked in the past or of being a current smoker and currently receiving anticoagulants was associated with a decreased risk from major haemorrhage.

Table 9-5 Univariable Fine and Gray models fits for arterial thrombotic events within 90 days (competing events: death and major haemorrhage), pooled over 10 MI data sets, and over possible trials.

	Fixed effect meta-analysis			Random effects meta-analysis			Heterogeneity		
Measure	**sHR**	**95%CI**	**P-value**	**sHR**	**95%CI**	**P-value**	***Q***	**df**	**P-value**
Age, per year	1.02	1.01-1.03	<0.0001	1.02	1.01-1.03	<0.0001	2.76	5	0.7375
Male	0.84	0.70-1.00	0.0558	0.83	0.65-1.08	0.1659	6.34	5	0.2745
Weight (kg)	0.99	0.98-1.00	0.0009	0.99	0.98-1.00	0.0285	7.8	5	0.1677
SBP (mmHg)	1.00	1.00-1.00	0.6573	1.00	1.00-1.00	0.6573	1.61	5	0.9005
DBP (mmHg)	1.00	0.99-1.00	0.4306	1.00	0.99-1.00	0.4306	1.64	5	0.8962
Atrial Fibrillation	1.55	1.28-1.89	<0.0001	1.55	1.28-1.89	<0.0001	2.54	5	0.7705
Hypertension	1.27	1.02-1.59	0.0325	1.27	1.02-1.59	0.0325	1.01	5	0.9618
Any vascular event	1.39	1.14-1.69	0.0012	1.39	1.14-1.69	0.0012	0.45	3	0.9304
Diabetes Mellitus	1.52	1.24-1.86	0.0001	1.52	1.19-1.94	0.0008	5.45	5	0.3637
Baseline NIHSS	1.02	1.00-1.04	0.0151	1.02	1.00-1.05	0.0736	7.84	5	0.1652
Abnormal hemisphere	1.12	0.93-1.35	0.2274	1.11	0.91-1.36	0.2891	5.17	5	0.3956
Abnormal ECG	1.97	1.59-2.45	<0.0001	1.97	1.59-2.45	<0.0001	3.87	5	0.5677
Antiplatelets									
Prior	1.57	1.30-1.90	<0.0001	1.57	1.30-1.90	<0.0001	2.52	4	0.6412
Current	0.97	0.78-1.21	0.8061	0.86	0.56-1.31	0.4732	3.77	3	0.2878
Anticoagulants									
Prior	1.60	1.17-2.17	0.0029	1.60	1.17-2.17	0.0029	0.97	2	0.6145
Current	1.16	0.93-1.46	0.1925	1.16	0.93-1.46	0.1925	2.68	3	0.4433
Prior or current smoker	1.08	0.89-1.32	0.4343	1.08	0.76-1.52	0.6783	6.04	4	0.1961

Table 9-6 Univariable Fine and Gray models fits for major haemorrhagic events within 90 days (competing events: death and arterial thromboses), pooled over 10 MI data sets, and over possible trials.

	Fixed effect meta-analysis			Random effects meta-analysis			Heterogeneity		
Measure	**sHR**	**95%CI**	**P-value**	**sHR**	**95%CI**	**P-value**	***Q***	**df**	**P-value**
Age, per year	1.02	1.02-1.03	<0.0001	1.02	1.02-1.03	<0.0001	3.65	6	0.7241
Male	1.10	0.91-1.34	0.3295	1.08	0.87-1.35	0.4852	6.50	6	0.3691
Weight (kg)	1.00	0.99-1.00	0.1392	1.00	0.99-1.00	0.1392	5.26	6	0.5105
SBP (mmHg)	1.00	1.00-1.00	0.6385	1.00	1.00-1.01	0.1871	10.12	6	0.1196
DBP (mmHg)	0.99	0.99-1.00	0.0574	1.00	0.99-1.01	0.7618	9.13	6	0.1663
Atrial Fibrillation	1.82	1.48-2.24	<0.0001	1.87	1.48-2.37	<0.0001	6.58	6	0.3610
Hypertension	1.16	0.92-1.47	0.2106	1.16	0.92-1.47	0.2155	6.05	6	0.4177
Any vascular event	1.06	0.86-1.31	0.5801	1.07	0.78-1.46	0.6715	7.48	5	0.1873
Diabetes Mellitus	1.15	0.92-1.44	0.2308	1.15	0.92-1.44	0.2308	3.76	5	0.5849
Baseline NIHSS	1.07	1.06-1.09	<0.0001	1.07	1.04-1.09	<0.0001	12.90	6	0.0447
Abnormal hemisphere	1.02	0.84-1.24	0.8248	1.02	0.84-1.24	0.8248	1.20	6	0.9771
Abnormal ECG	1.52	1.22-1.91	0.0002	1.52	1.22-1.91	0.0001	1.10	5	0.9544
Antiplatelets									
Prior	1.37	1.12-1.68	0.0026	1.37	1.12-1.68	0.0026	2.58	6	0.8597
Current	0.46	0.36-0.58	<0.0001	0.46	0.36-0.58	<0.0001	2.58	4	0.6311
Anticoagulants									
Prior	1.90	1.39-2.59	0.0001	2.05	1.31-3.21	0.0017	6.09	4	0.1924
Current	0.96	0.75-1.23	0.7449	1.00	0.66-1.50	0.9918	5.91	4	0.2060
Prior or current smoker	0.80	0.64-1.00	0.0460	0.80	0.64-1.00	0.0460	1.66	4	0.7986

9.5.4.2 Multivariable prediction models for adverse events

Multivariable FG models were developed using data from *Trial 1* in the VISTA extract. This was the largest of the seven trials and the most complete (see Table 9-2 and Table 9-3). The remaining trials were used for testing model performance.

Multiple Fractional Polynomials' (MFP) were fit for all continuous measurements which included: patient age; weight; systolic BP; and NIHSS. The stability of the imputations for the continuous variables in *Trial 1* is illustrated in Figure 9-4 with each of the 10 obtained imputed sample distributions overlain in the same plot. There was no discernible variability in the covariate distributions across the ten imputed datasets. The *majority* MFP approach supported the same transformation for each of the 10 imputed datasets. Only one non-linear transformation was chosen for the *major haemorrhage* FG model which was a natural log transformation for NIHSS. This was chosen consistently across each of the imputed sets – all other continuous variables entered the FG competing risk models as linear. The NIHS Score ranges from 0 to 42 increasing in unit increments. Since a natural zero exists within the NIHSS range it was necessary to shift all values up by a single unit. After selection, these transformations were all fit within the FG model framework using a single imputed dataset. The final multivariable FG models are presented in Table 9-7. Under the 10EPV rule of thumb there were a total of 31 candidate degrees of freedom for the arterial thrombosis model and 24 candidate degrees of freedom for the major haemorrhage model. The assumption of proportionality was assessed by visually inspecting residual plots. There was no indication from these plots that the covariate-outcome effects depended on time (see section 9.8 Appendix B on page 340). Risk of arterial thrombotic events was associated with: increasing age (per 10 years); being a current or previous smoker; having an abnormal ECG; diabetic; a prior history of taking antiplatelets; a prior history of taking anticoagulants; and a concurrent record of antiplatelets. Similarly, increasing age (per 10 years); increasing NIHSS; being male; and concurrent antiplatelets were all associated with an increased risk of major haemorrhage.

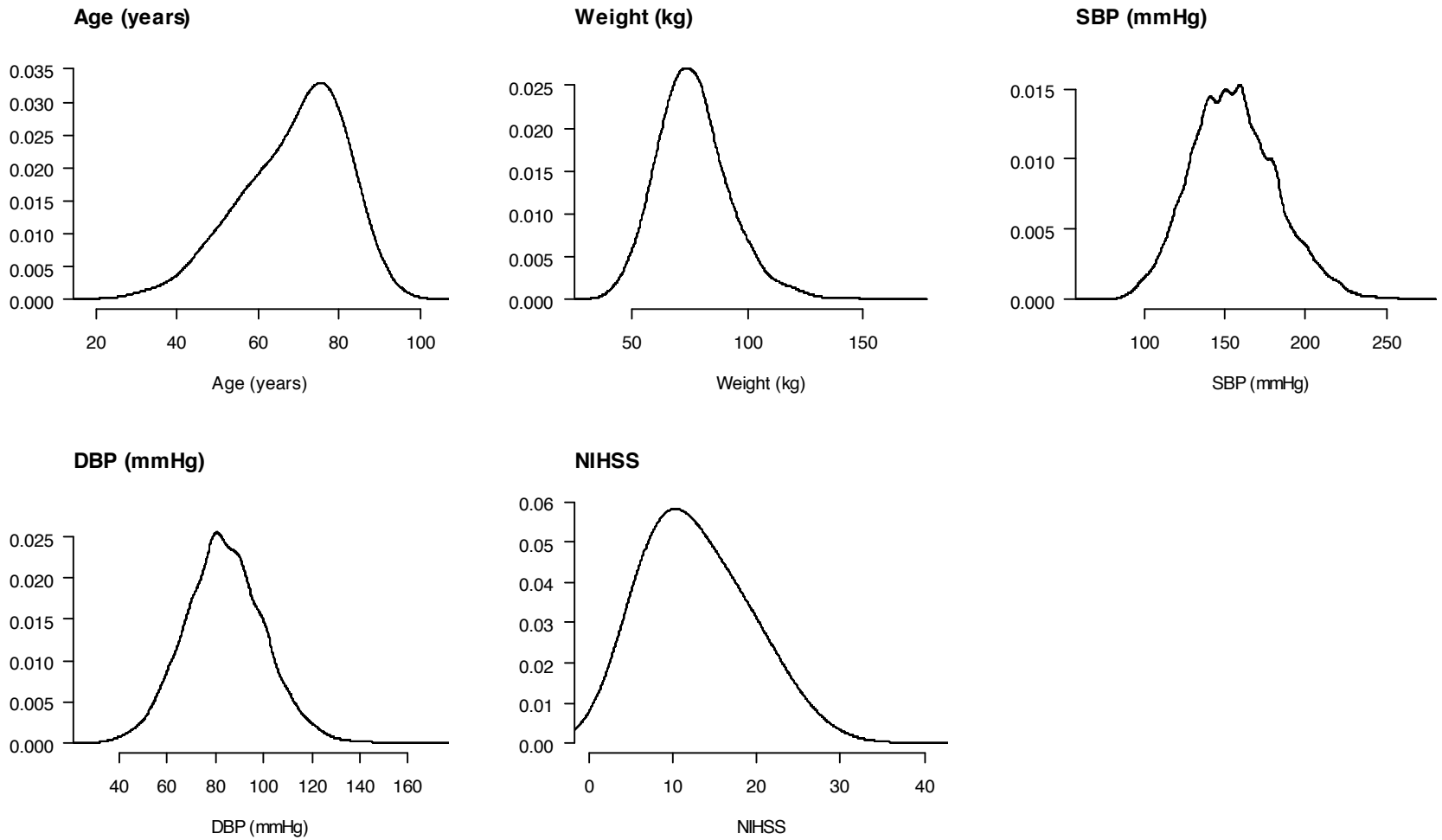


Figure 9-4 Kernel density plots for each of the continuous variable in the VISTA dataset with each of the imputed sets superimposed.

Table 9-7 Fine and Gray regression models for the prediction of: (i) arterial thrombosis and (ii) major haemorrhage.

	Arterial thrombosis competing risk model			Major haemorrhage competing risk model		
Variable	**Coefficient (SE)**	**sHR (95% CI)**	**P-value**	**Coefficient (SE)**	**sHR (95% CI)**	**P-value**
Age, per 10 years[a]	0.133 (0.056)	1.14 (1.02-1.27)	0.0173	0.159 (0.064)	1.17 (1.03-1.33)	0.0125
SBP, per 10mmHg[a]	-0.006 (0.021)	0.99 (0.96-1.04)	0.7883	-0.031 (0.024)	0.97 (0.92-1.02)	0.1913
Weight, per 10kg[a]	-0.073 (0.045)	0.93 (0.85-1.02)	0.1043	-0.008 (0.046)	0.99 (0.91-1.09)	0.8567
NIHSS[a]						
Linear	-0.005 (0.011)	1.00 (0.97-1.02)	0.6634	-	-	-
Log (*x* + 1)	-	-	-	1.330 (0.185)	3.78 (2.63-5.43)	<0.0001
Gender, male	-0.143 (0.130)	0.87 (0.67-1.12)	0.2687	0.412 (0.144)	1.51 (1.14-2.00)	0.0042
Current or previous smoker	0.258 (0.124)	1.29 (1.02-1.65)	0.0372	-0.155 (0.151)	0.86 (0.64-1.15)	0.3031
ECG, abnormal scan	0.484 (0.150)	1.62 (1.21-2.18)	0.0012	-0.011 (0.169)	0.99 (0.71-1.38)	0.9488
Hemisphere, right vs. left side	0.164 (0.117)	1.18 (0.94-1.48)	0.1601	0.216 (0.134)	1.24 (0.95-1.61)	0.1068
History of diabetes	0.396 (0.130)	1.49 (1.15-1.92)	0.0024	0.037 (0.160)	1.04 (0.76-1.42)	0.8176
History of hypertension	0.021 (0.145)	1.02 (0.77-1.36)	0.8876	0.077 (0.162)	1.08 (0.79-1.48)	0.6353
History of atrial fibrillation	0.207 (0.135)	1.23 (0.94-1.60)	0.1252	0.025 (0.166)	1.03 (0.74-1.42)	0.8816
History of any vascular event[b]	0.042 (0.124)	1.04 (0.82-1.33)	0.7306	-0.143 (0.142)	0.87 (0.66-1.15)	0.3159
Prior medication						
Any antiplatelet[c]	0.368 (0.124)	1.44 (1.13-1.84)	0.0031	0.136 (0.144)	1.15 (0.86-1.52)	0.3453
Any anticoagulant[d]	0.437 (0.194)	1.55 (1.06-2.26)	0.0239	0.294 (0.232)	1.34 (0.85-2.12)	0.2063
Concurrent medication						
Any antiplatelet[c]	0.296 (0.123)	1.35 (1.06-1.71)	0.0156	-0.641 (0.147)	0.53 (0.39-0.70)	<0.0001
Any anticoagulant[d]	0.137 (0.128)	1.15 (0.89-1.47)	0.2849	-0.161 (0.152)	0.85 (0.63-1.15)	0.2901

a Functional form of continuous variables selected using a multiple fractional polynomial routine chosen at P-value <0.157

b Any previous strokes, TIAs or MIs

c Any used of aspirin, clopidogrel, dipyridamole or ticlopidine

d Any use of heparin or warfarin

9.5.4.3 Model performance

Model performance was evaluated in six of the seven available trials, five of which offered an assessment of the external validity of the Fine and Gray competing event model fits. For each of the trials missing data were imputed using multiple imputation methods generating 10 completed datasets, therefore each of the performance measures provided per trial in Table 9-8 (i.e., the *c*-statistic and calibration slope) corresponds to a pooled estimate over 10 imputed datasets using Rubin's rules (Vergouwe et al., 2010). Across the five external trial datasets, the prediction model for arterial thrombosis achieved *c*-statistics which ranged from as low as 0.50 to as high as 0.67, suggesting no discrimination to moderate discrimination depending on which trial was considered. The corresponding pooled random effects estimate of those evaluation *c*-statistics was 0.59 (95%CI 0.52 to 0.66) with a 95%PI of 0.37 to 0.81. Similarly, for major haemorrhage the *c*-statistics ranged from as low as 0.57 to as high as 0.74, suggesting poor to good discrimination depending on which trial was considered. The pooled random effects estimate was 0.64 (95%CI 0.58 to 0.70) with a 95%PI of 0.46 to 0.82.

Calibration was not consistent across trials indicating that in addition to potential over-fitting within the development dataset there were real differences in the risk profiles of patients across the individual trials. The incidence of adverse events differed from trial to trial and in all but one of the five external datasets the Fine and Gray models over-predicted the observed 90 day risk from both events (Table 9-8).

The calibration slope gives an indication as to how well the linear predictor generalises from the development cohort to the evaluation cohort and thus summarises whether the relative risks were well specified by the proposed fit (van Houwelingen, 2000). In some trials the models were a good fit but in others they were very poor. The 95% prediction intervals illustrate the extent of this suggesting that based on these data 95% of any estimates obtained for future trials the effect sizes amongst these models will could range from too large right through to too small without updating (Steyerberg, 2009). More detail regarding calibration appears in the appendix at the end of this chapter (see section 9.9 Appendix C on page 345).

Table 9-8 Model performance of Fine and Gray regression models within six trial datasets pooled over 10 imputed datasets.

		Discrimination			Predicted 90-day risk[c]		Calibration[b]		
Trial	**Model**	***c*-index**	**95% CI[a]**	**95% PI**	**Estimated**	**Observed**	**Slope**	**95% CI**	**95% PI**
1 (Development-set)	Thrombosis	0.64	0.61 to 0.67	-	6.29%	6.29%	1.00	0.79 to 1.21	-
	Haemorrhage	0.71	0.67 to 0.74	-	4.76%	4.77%	1.00	0.82 to 1.16	-
2	Thrombosis	0.50	0.37 to 0.64	-	6.24%	2.79%	0.00	-0.80 to 0.80	-
	Haemorrhage	0.74	0.63 to 0.85	-	7.56%	4.01%	1.47	0.44 to 2.51	-
4	Thrombosis	0.67	0.60 to 0.73	-	5.24%	4.11%	1.16	0.67 to 1.66	-
	Haemorrhage	0.61	0.53 to 0.69	-	6.95%	3.80%	0.53	0.11 to 0.96	-
5	Thrombosis	0.65	0.58 to 0.72	-	4.75%	3.88%	1.02	0.43 to 1.61	-
	Haemorrhage	0.57	0.49 to 0.65	-	6.86%	3.03%	0.34	-0.04 to 0.71	-
6	Thrombosis	0.50	0.37 to 0.64	-	4.07%	4.18%	-0.06	-0.96 to 0.84	-
	Haemorrhage	0.62	0.52 to 0.72	-	6.16%	7.57%	0.40	0.02 to 0.78	-
7	Thrombosis	0.53	0.38 to 0.68	-	4.17%	3.49%	0.38	-0.82 to 1.58	-
	Haemorrhage	0.73	0.56 to 0.89	-	4.40%	2.99%	1.39	0.25 to 2.52	-
Pooled RE model estimate[d]	Thrombosis	0.59	0.52 to 0.66	0.37 to 0.81	-	-	0.59	0.06 to 1.13	-1.12 to 2.31
	Haemorrhage	0.64	0.58 to 0.70	0.46 to 0.82	-	-	0.57	0.25 to 0.88	-0.29 to 1.42

a Boot strap 95% CIs are provided for each of the individual estimates per trial (over 1000 replications). Estimated SEs were used to obtain the pooled estimates
b Estimated risk obtained from the mean predicted risk from the FG models. Observed risk obtained from cumulative incidence function
c Calibration slope estimated via a Fine and Gray model with the linear predictor produced for the corresponding trial as the only variate
d Pooled estimates exclude metrics obtained from the development set which will be too optimistic

9.5.5 Mortality rate stratified by predicted risk

The predicted risk from arterial thrombosis and major haemorrhage can be obtained using the two multivariable FG models presented in Table 9-7 at any time-point within the 90 day follow-up period. Predicted risks were evaluated using the `predict.crr` function in the `cmprsk` R library which combines a Breslow-type estimate of the underlying hazard with the linear predictor for each patient across the period of follow-up (Gray, 2013).

Each patient is given a predicted 90 day cumulative risk of major haemorrhage and predicted 90 day cumulative risk of arterial thrombosis (Figure 9-5). Excluding the development set, all those patients with intermediate events in follow-up are plotted irrespective of trial. When restricting attention to the margins in the left-hand side plot, it is seen that those with observed major haemorrhages and those with observed arterial thromboses in follow-up have similarly distributed predicted risks with little distinction between the two. The right-hand side plot shows a simple random sample (SRS) of 800 patients who were either dead or alive and event free for illustrative purposes. Again, looking at the marginal distributions (calculated on all patients who were either alive and event free or dead) for predicted 90 day cumulative risk of arterial thrombosis risk and of major haemorrhage it can be seen that the predicted 90 day cumulative risk of major haemorrhage is far better at separating those who are dead from those who are alive and event free than the arterial thrombosis model.

Finally, it is noted that there was no indication that particular adverse event types were identifiable using the predictions obtained from these models.

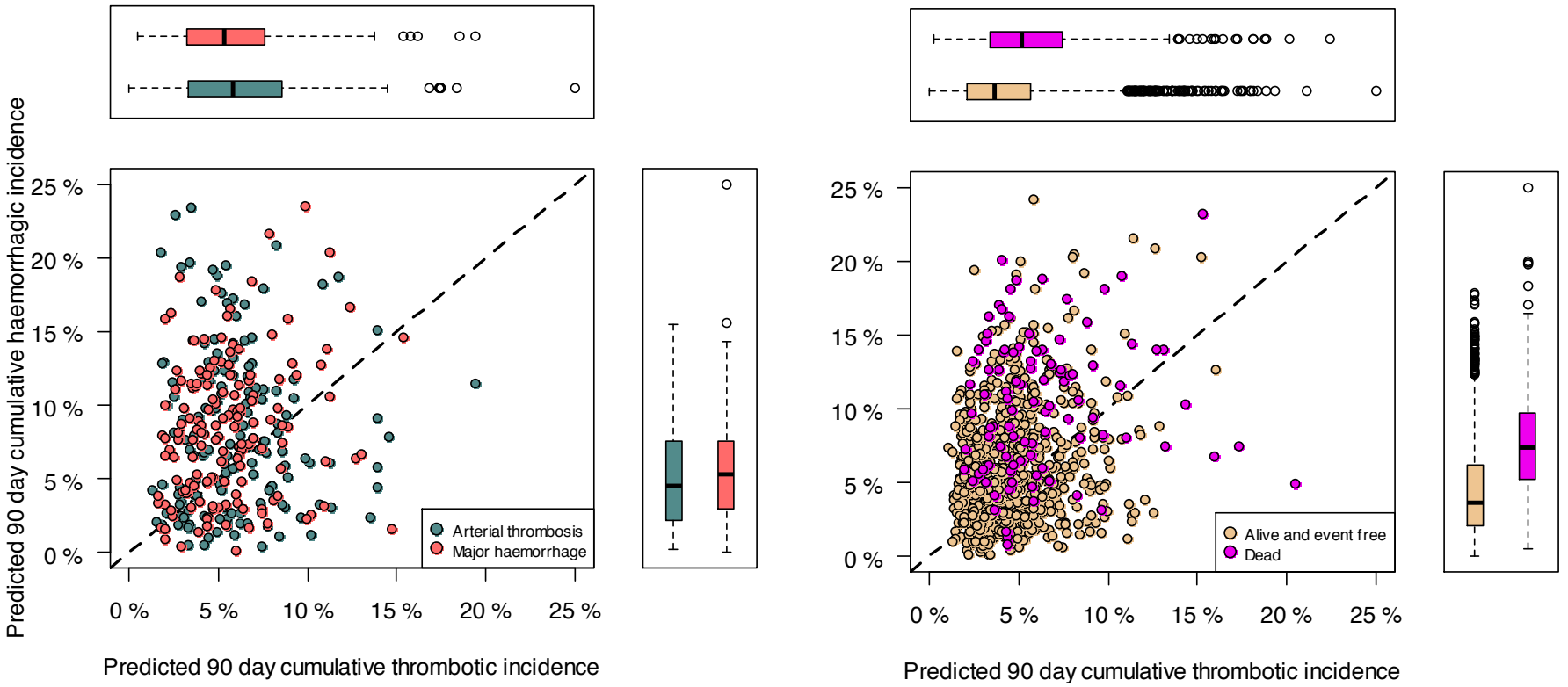


Figure 9-5 Predicted risk of arterial thrombosis vs. predicted risk of major haemorrhage. The left hand plot shows all recorded major haemorrhagic and arterial thrombotic events, whilst the right hand side displays a SRS of 800 patients who were either alive or dead. Marginal box plots are based on summary statistics from all patients.

Using the median predicted 7-day risk estimated within the development dataset (*Trial 1*) it is possible to specify a categorisation of predicted adverse event risk creating four distinct strata of low versus high risk of major haemorrhage and arterial thrombosis, i.e.: (i) low risk of major haemorrhage and low risk of arterial thrombosis; (ii) low risk of major haemorrhage and high risk of arterial thrombosis; (iii) high risk of major haemorrhage and low risk of arterial thrombosis; and finally (iv) high risk of major haemorrhage and high risk of arterial thrombosis. Within each of the identifiable risk stratum the cumulative risk of mortality can be estimated across the five evaluation trial datasets which could then be pooled.

In Figure 9-6 each grey box represents the random effects point estimate with the corresponding 95%CI and assessment of trial heterogeneity (i.e., *Q*-statistic, degrees of freedom (*df*) and *I*-squared value). The black diamond shows the overall pooled estimate. Contrasting the low and high risk strata for one event type whilst holding the other event type constant is a useful way of exploring which of the predicted risk groups has the largest impact on the rate of mortality. There was a considerable increase in the rate of mortality when moving from low to high risk from a major haemorrhage whilst holding the risk of arterial thrombosis constant (low or high)

The individually obtained estimates had a large degree of between trial heterogeneity with large *I*-squared values as high as 77% in some strata. However, this was anticipated since the incidence of the individual outcomes split by trial (see Table 9-2and Table 9-3) ranged between 6% and 18%. It is noted that in another meta-analysis which combined the cumulative risks obtained from various studies over time a similar high degree of between study heterogeneity was identified (Mohan et al., 2011).

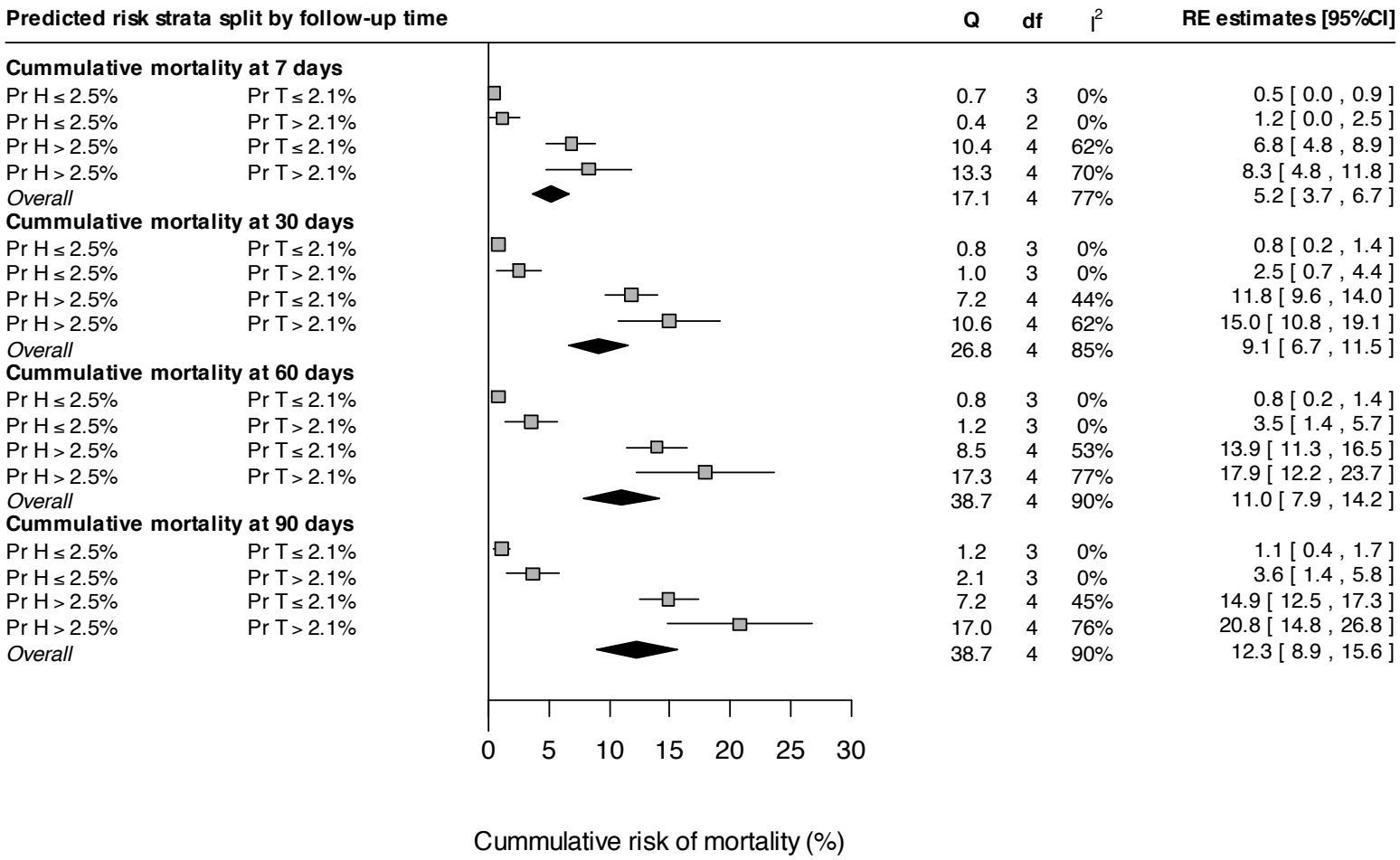


Figure 9-6 Pooled K-M estimates split by predicted 7 day risk of major haemorrhage and arterial thrombosis (for details see text).

9.6 Discussion

Patient survival following an acute ischaemic stroke is heterogeneous. This variability is explained at least in part by various measurable characteristics regarding: the severity of the stroke; patient's age; as well as their underlying medical history (Weimar et al., 2004). Intervening events that occur over the course of recovery also impact the rate of mortality; although attributing a single constant hazard ratio to each adverse event type may not be plausible, indeed, the impact on mortality may interact with time. This interaction may though be biologically plausible as those who suffer an adverse event late in follow-up will have had longer to recover and therefore be in a better position to overcome an adverse event. Whereas those that suffer an adverse event early on in follow-up will be in a poorer condition and therefore the impact could indeed cause a greater increase in the hazard of mortality.

Clinical prediction models developed for predicting patient risk from major haemorrhage and arterial thrombosis over a 90-day follow-up period performed moderately well in evaluation. However, this was highly dependent upon the dataset used. The prediction model developed to predict major haemorrhage after acute ischaemic stroke achieved good discrimination for the most part though was poorly calibrated in different trials. It is important to recognise that calibration is a joint property of both *model* and *data* (Vickers and Cronin, 2010), though it was not possible to investigate whether poor calibration could be attributed to particular trial level characteristics and therefore these models require further assessment in data were aspects of design etc. is known.

The data analysed in this chapter are made up of a highly selected sample of stroke patients who are unlikely to represent patients typically seen during standard clinical practice. These results should therefore be interpreted with caution. They could, however, play a role in the recruitment of patients to future randomised trials for the purpose of enriching the cohort of patients.

One of the strengths of this analysis is that it illustrates the application of novel techniques for the development, evaluation, and application of clinical prediction

models in the presence of competing events. However, further work is warranted. For instance, whilst a competing risk framework enables the modelling of covariates for the risk of the primary event of interest in the presence of a competing event, a more complex multistate system is more plausible in practice. Under such a system, patients can either: suffer a single adverse event type after which they will either survive or die; alternatively they may suffer multiple event types again following which they will either survive or die. By modelling each of the cause-specific hazard pathways it would be possible to explore how patient risk profiles differ according to common adverse events during recovery. A very large registry of routinely collected data would be needed to ensure a sufficient number of event types moving from one state to the next as well as a set of results that could be generalised to clinical practice.

9.7 Appendix A: Missing data

Missing data were explored per trial using hierarchical cluster plots each presented alongside a simple dot-plot summarising the number of missing variables per observation (see Figure 9-7 to Figure 9-13 below). The most common *jointly* missing variables were blood pressure (systolic and diastolic), in *Trials 4* and *5* the same variables were jointly missing: previous stroke; myocardial infarction; diabetes mellitus; hypertension; any (previous or current) smoking status; and atrial fibrillation. Similarly, in *Trial 7* those variables that were commonly jointly missing were: myocardial infarction; hypertension; atrial fibrillation; and previous TIA.

The proportion of joint missingness was smallest amongst *Trials 1*, *2* and *6*. From Trials 1 through to 7, the proportion of missingness varied with: 6%, 17%, 11%, 11%, 20%, 10% and 42% missing respectively (note that those completely missing variables are excluded from this assessment). The trial with the fewest complete data was therefore *Trial 7* with only 58% of patients with no missing values and 38% of those in this trial with 4 missing values or more.

It is noted that *Trial 3* had the greatest number variables with jointly missing entries, which included all but the following variables: previous or current treatment with antiplatelets or anticoagulants; age; and gender

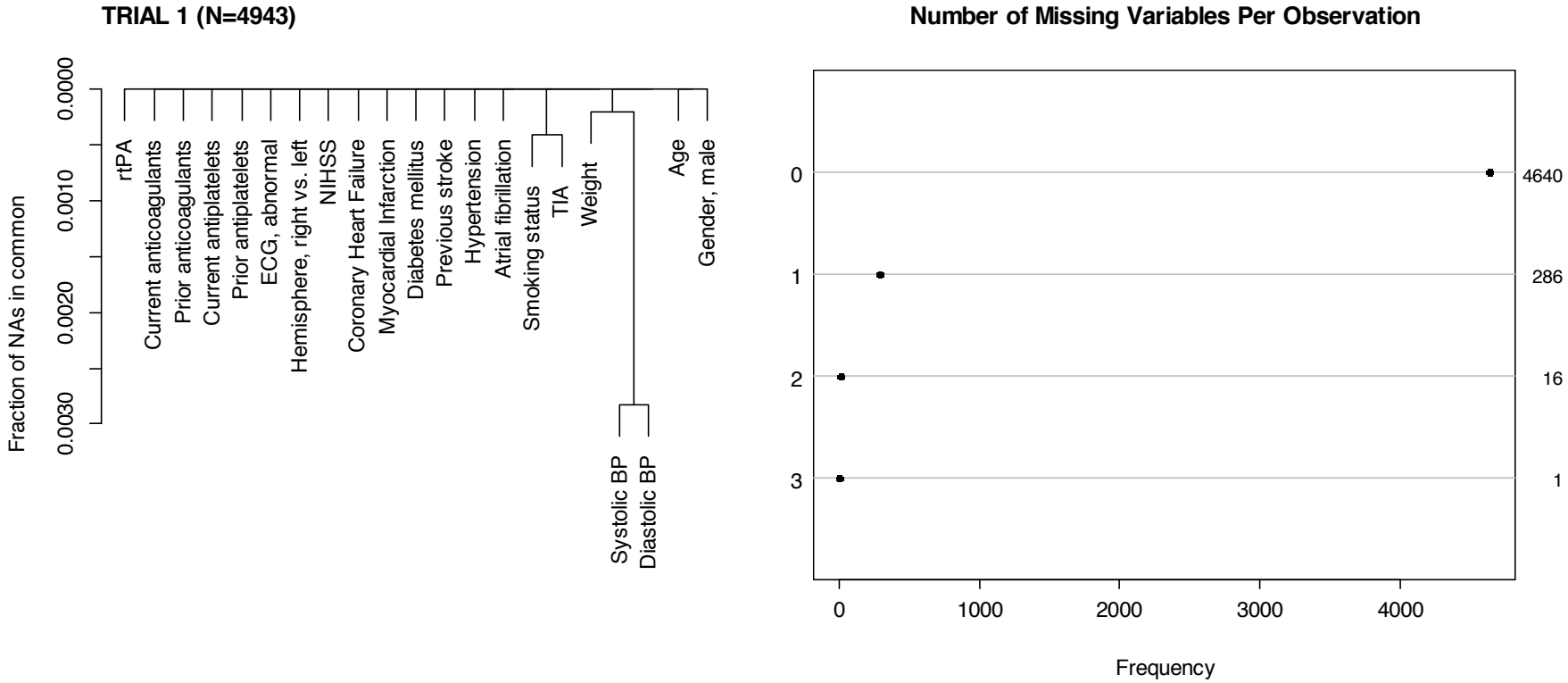


Figure 9-7 Combinations of missing values in Trial 1; a hierarchical cluster analysis of combined missingness and frequency of NAs per observation.

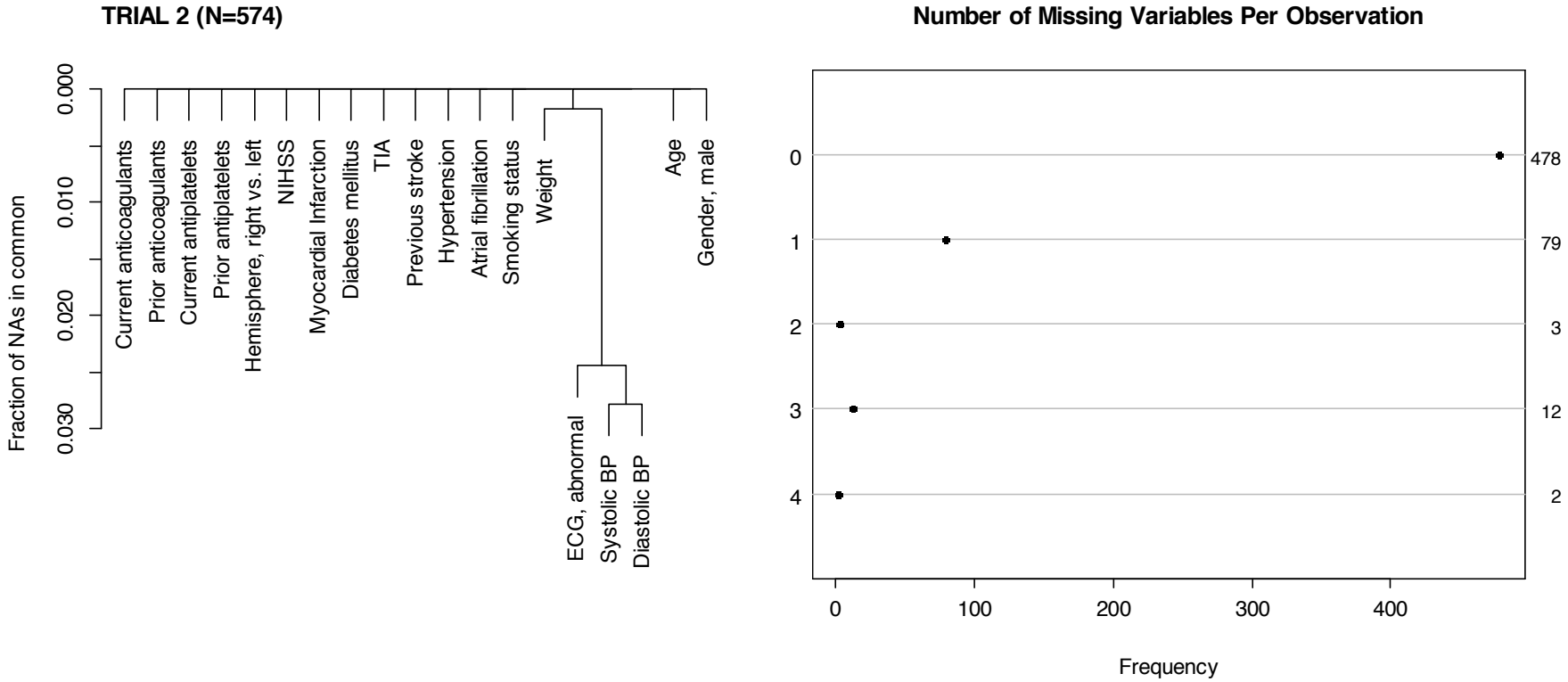


Figure 9-8 Combinations of missing values in Trial 2; a hierarchical cluster analysis of combined missingness and frequency of NAs per observation.

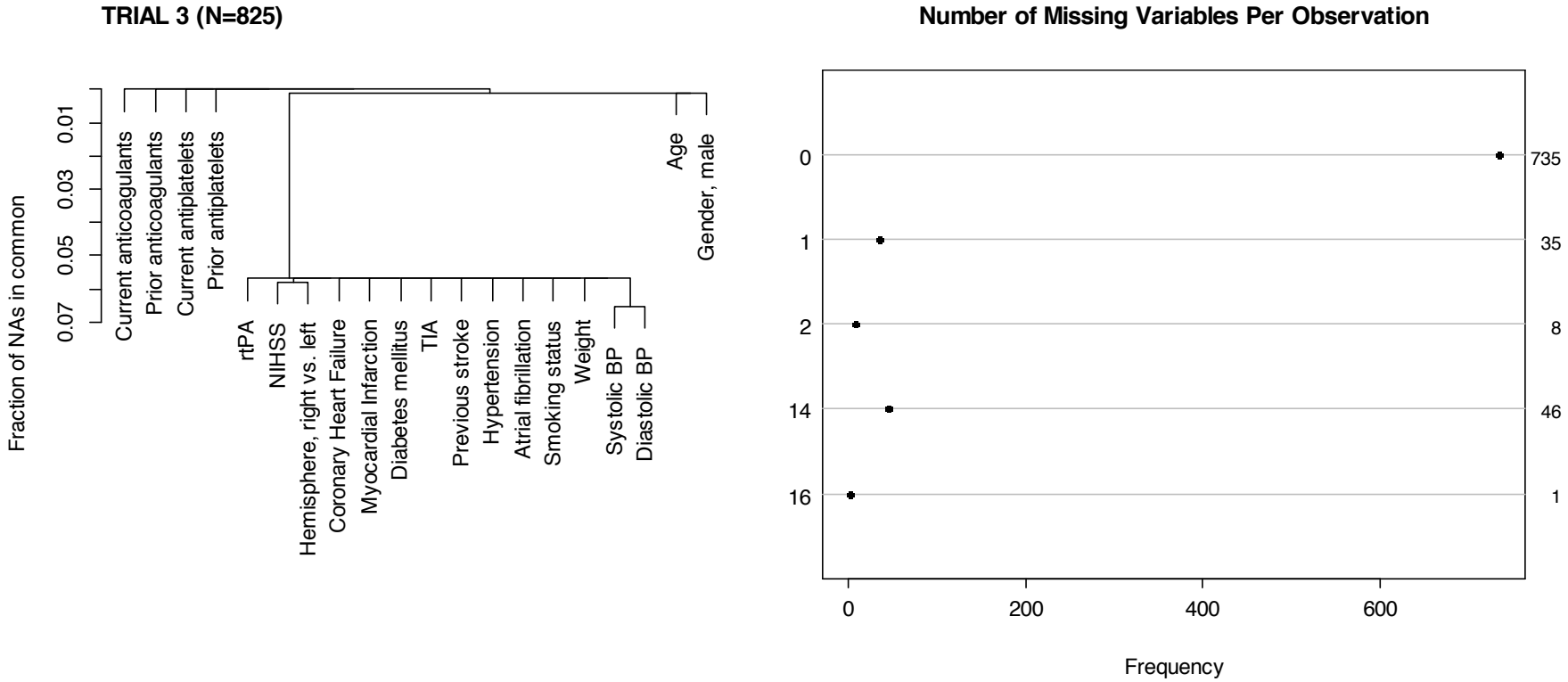


Figure 9-9 Combinations of missing values in Trial 3; a hierarchical cluster analysis of combined missingness and frequency of NAs per observation.

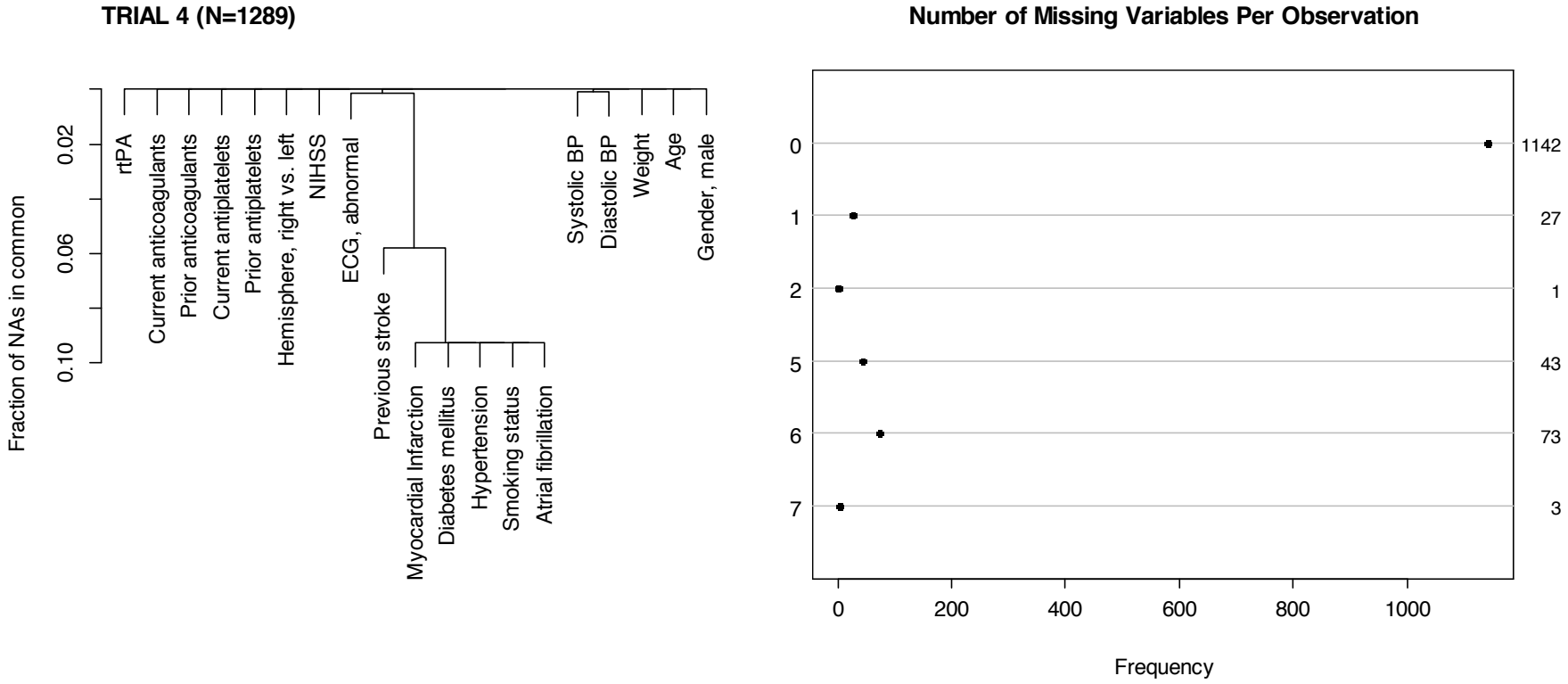


Figure 9-10 Combinations of missing values in Trial 4; a hierarchical cluster analysis of combined missingness and frequency of NAs per observation.

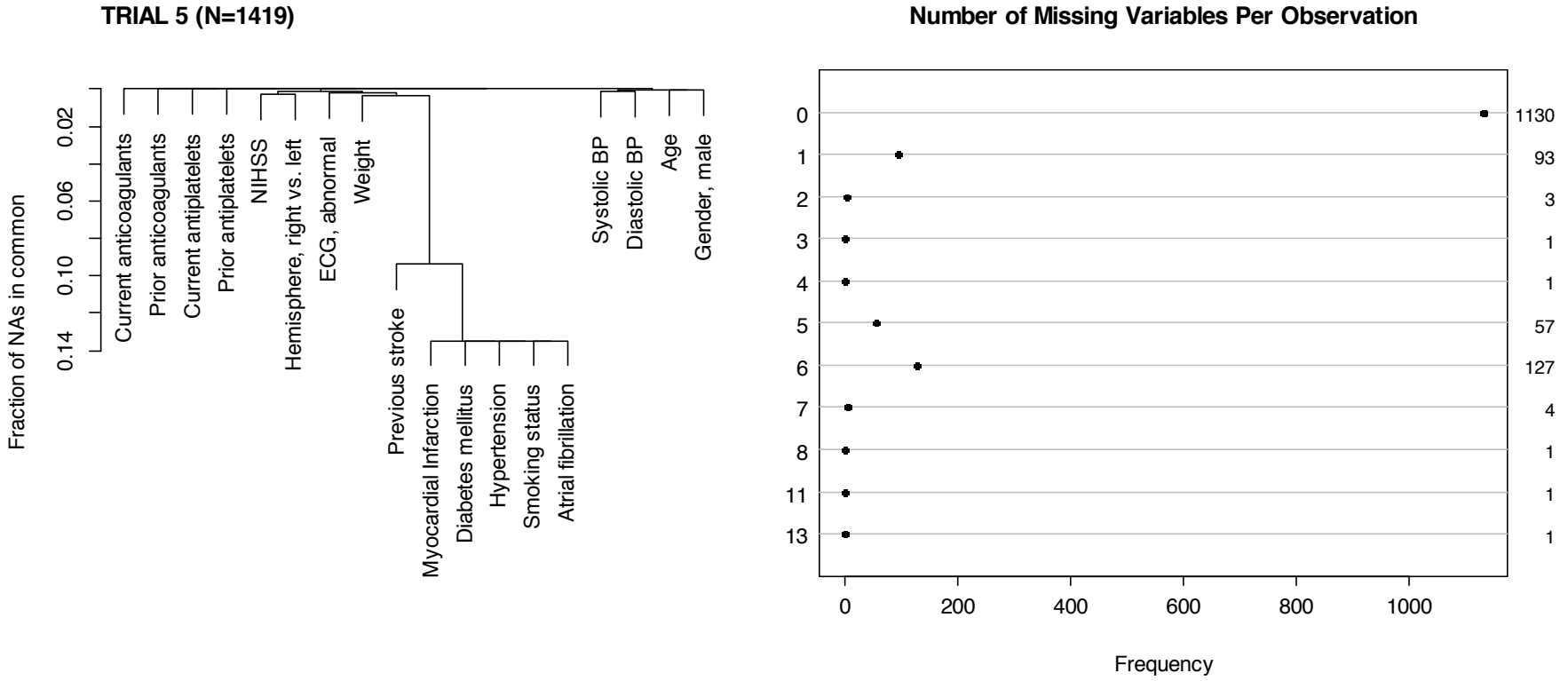


Figure 9-11 Combinations of missing values in Trial 5; a hierarchical cluster analysis of combined missingness and frequency of NAs per observation.

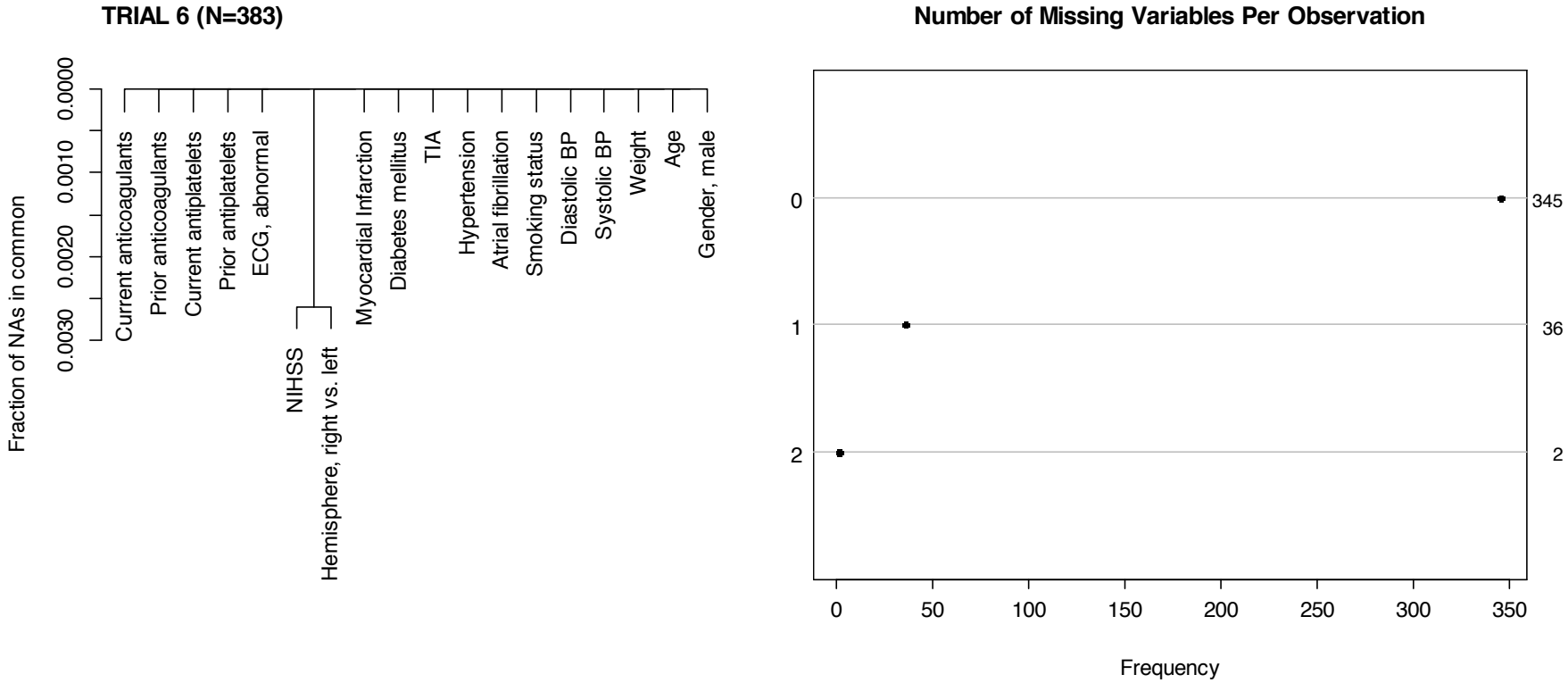


Figure 9-12 Combinations of missing values in Trial 6; a hierarchical cluster analysis of combined missingness and frequency of NAs per observation.

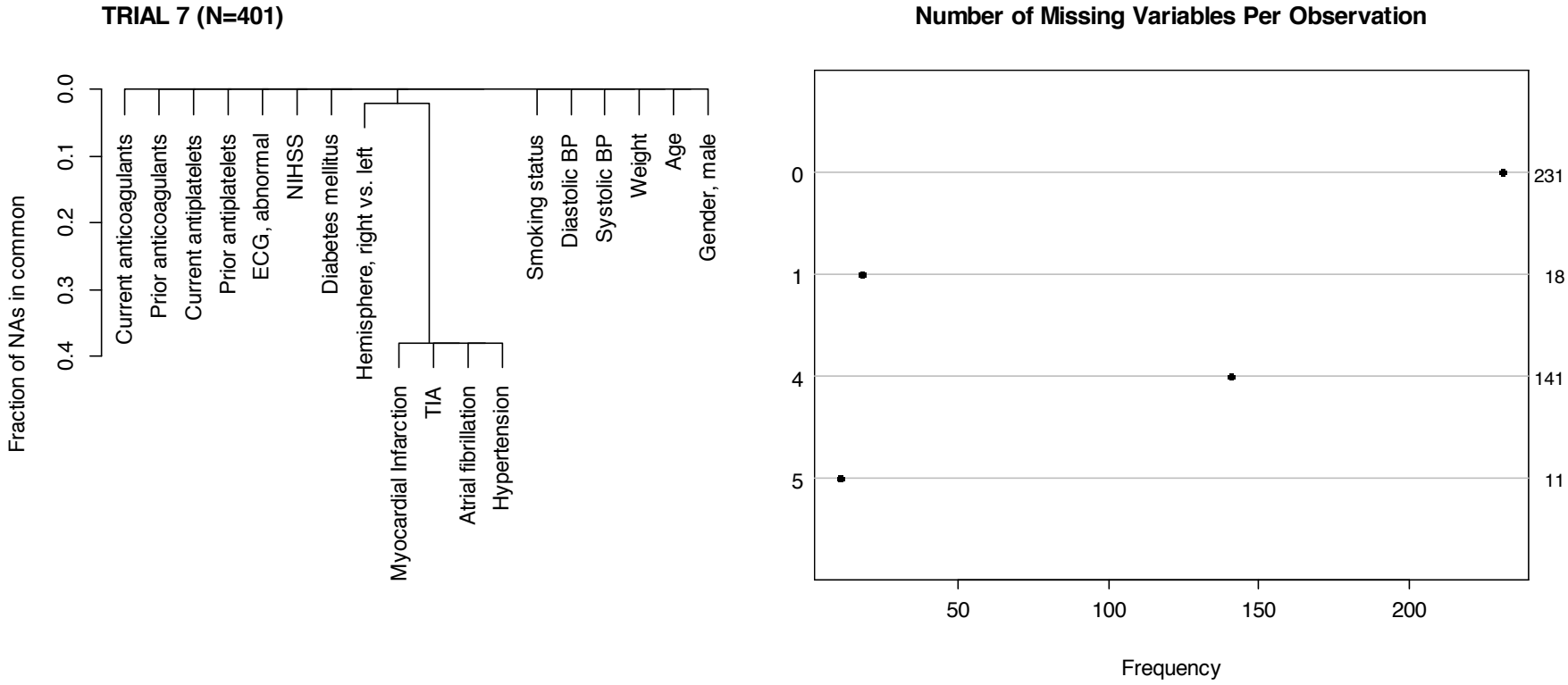


Figure 9-13 Combinations of missing values in Trial 7; a hierarchical cluster analysis of combined missingness and frequency of NAs per observation.

9.8 Appendix B: residual plots for Fine and Gray models

The plots below are used to assess whether or not the required proportionality assumption holds in the case of the two developed Fine and Gray models. The corresponding Schoenfeld *type* residuals were obtained from the `crr` function implemented in R (Gray, 2013). Under the proportionality assumption a plot of the residuals against the unique failure times should illustrate a random scatter of points with no discernable shape, i.e., no significant trends against time. A non-parametric smoothing function can be used to help guide the eye. In the residual plots that follow a lowess smoother was used to highlight the general trend of the scatter of points.

Figure 9-14 and Figure 9-15 show the residual plots for the covariate effects in the arterial thrombosis model. With no strong departures from a random scatter, the proportionality assumption looks as if it holds for these data. Similarly Figure 9-16 and Figure 9-17 show the residual plots for the covariate effects in the major haemorrhage model. Again there is no strong departure from a random scatter in any of the residual plots suggesting that the effects are not time-dependent.

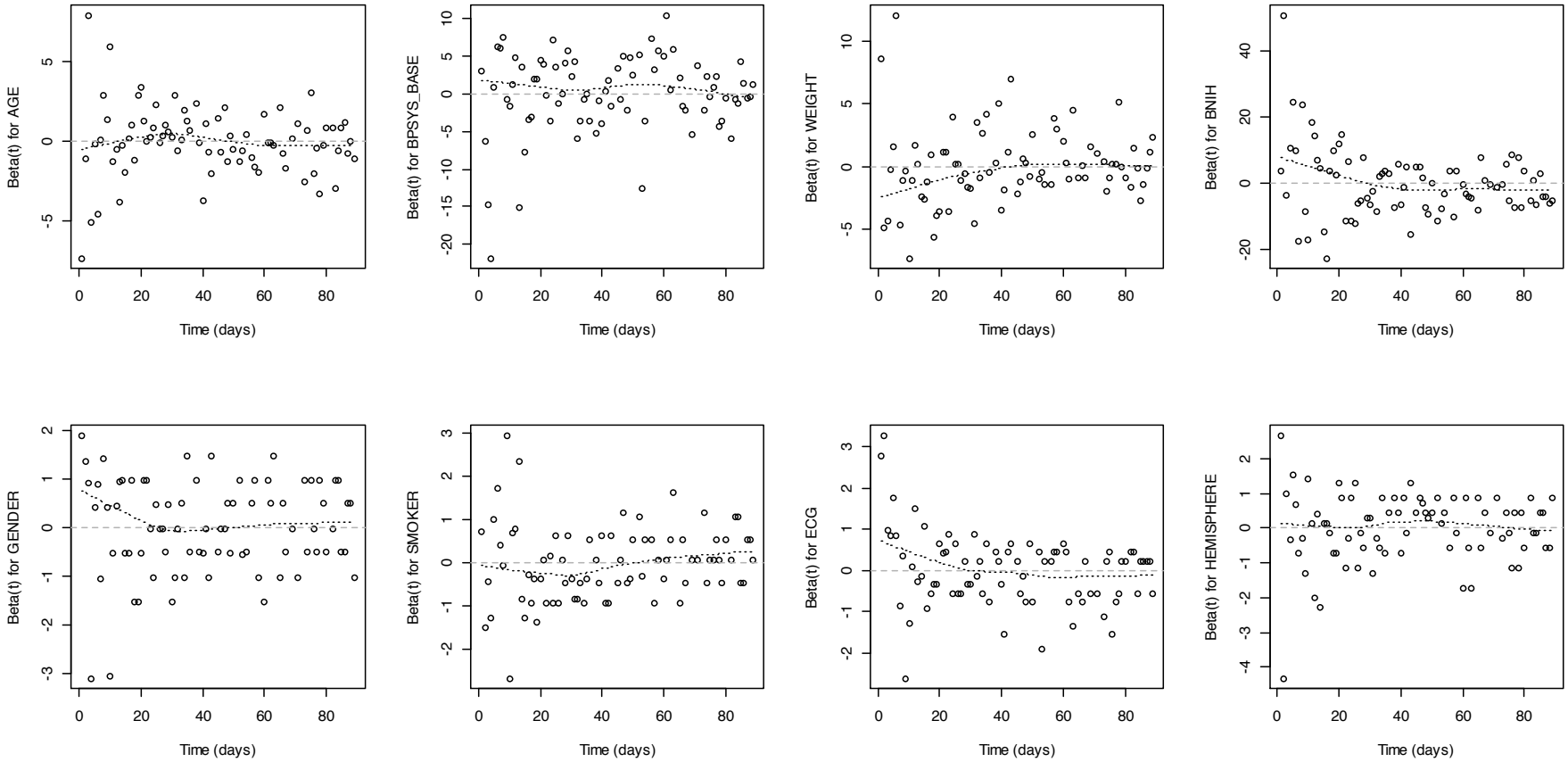


Figure 9-14 (i) Schoenfeld type residual plots for Fine and Gray competing risk model for the prediction of arterial thrombosis (continued in Figure 9-15).

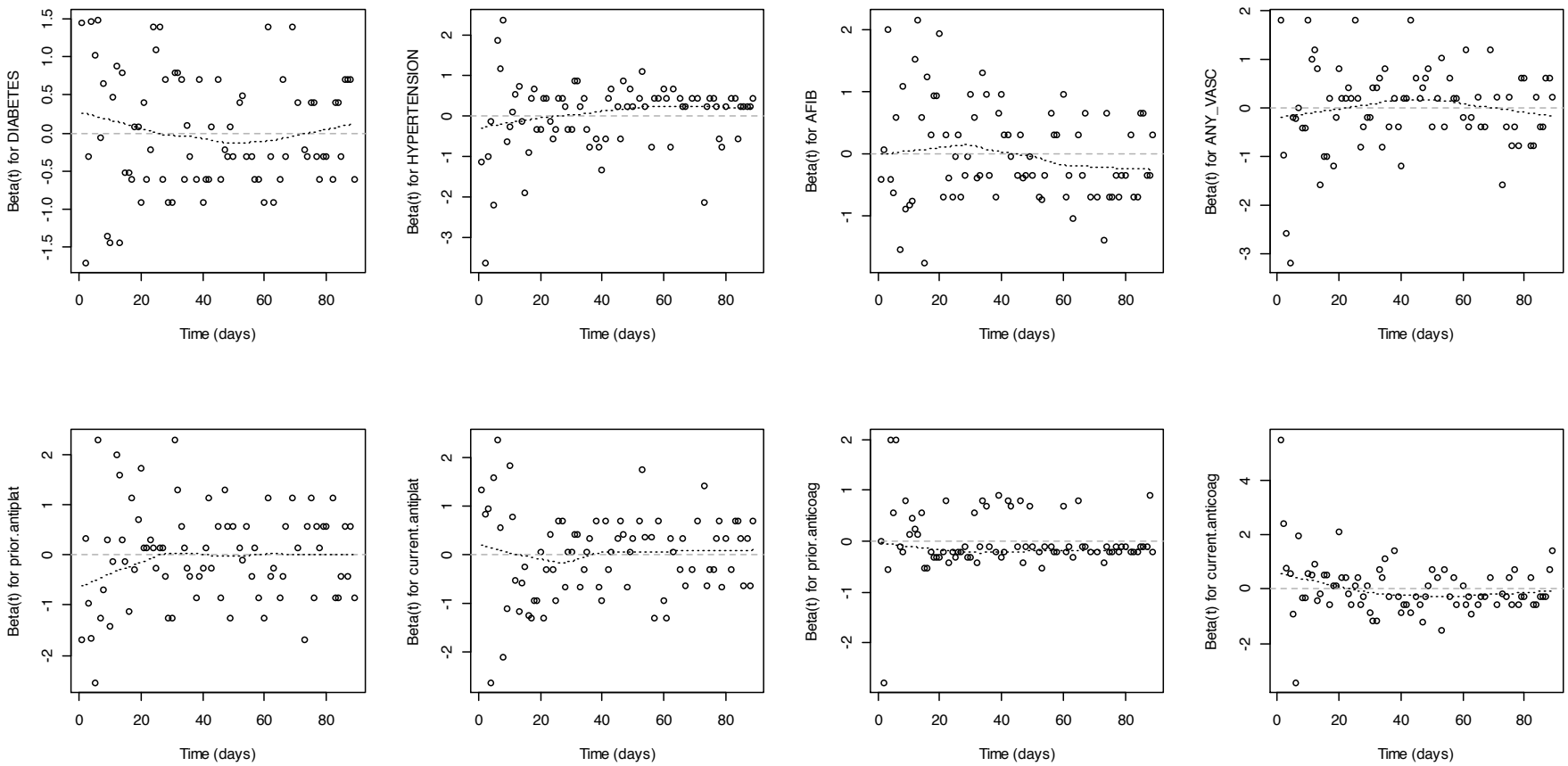


Figure 9-15 (ii) Schoenfeld type residual plots for Fine and Gray competing risk model for the prediction of arterial thrombosis.

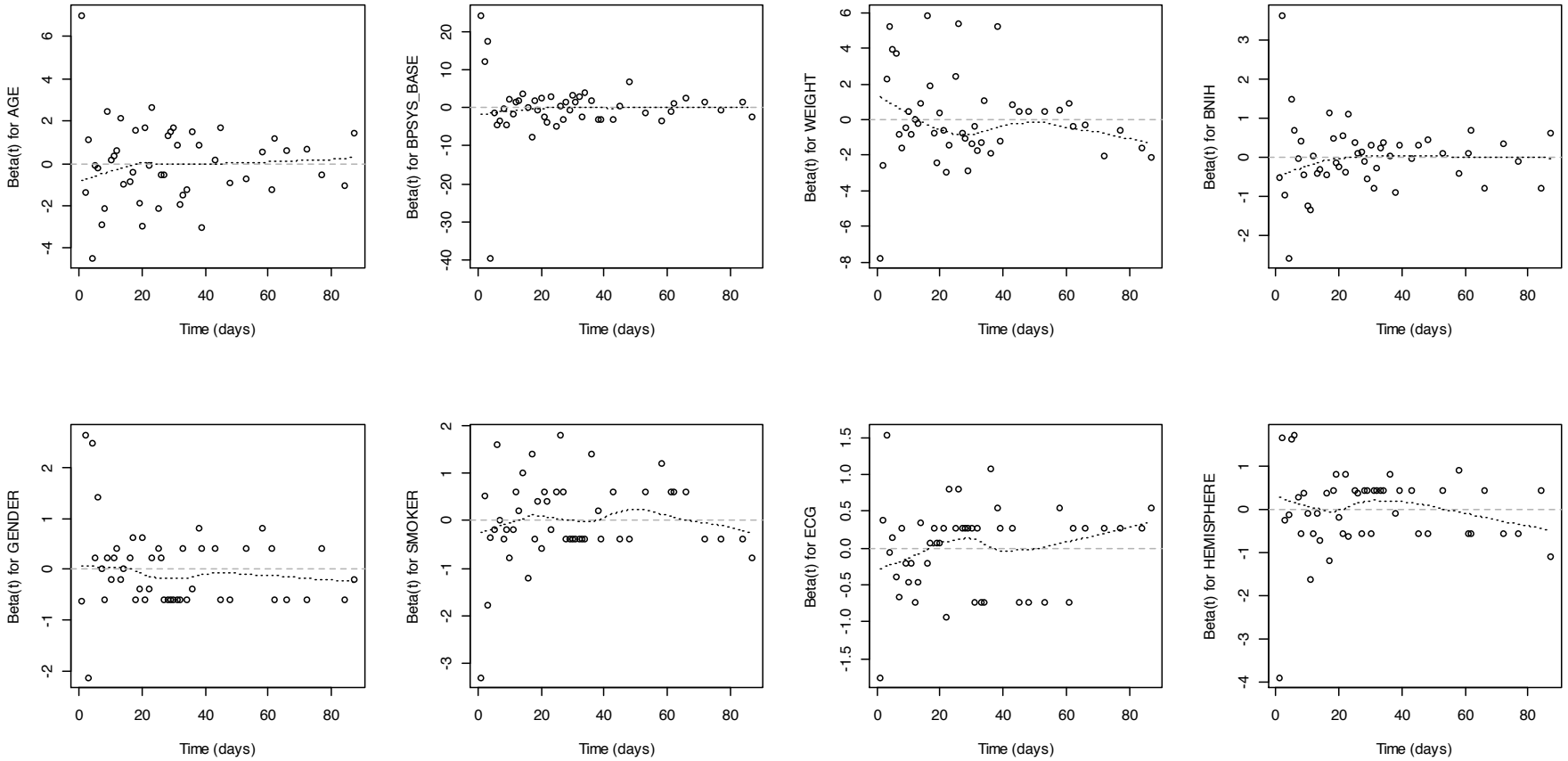


Figure 9-16 (i) Schoenfeld type residual plots for Fine and Gray competing risk model for the prediction of major haemorrhage (continued in Figure 9-17).

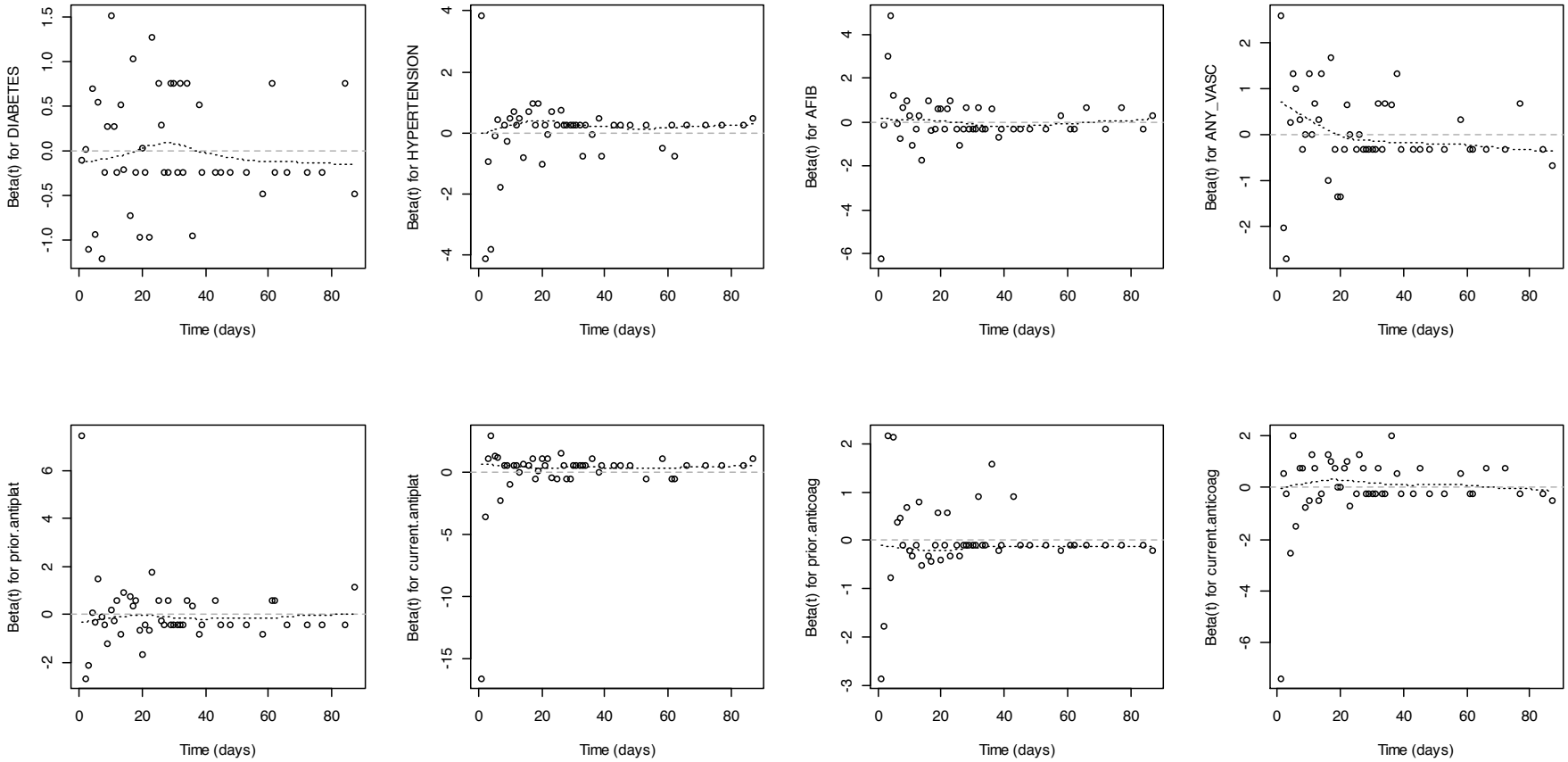


Figure 9-17 (ii) Schoenfeld type residual plots for Fine and Gray competing risk model for the prediction of major haemorrhage.

9.9 Appendix C: Calibration of models

Calibration plots are provided for each of the trial datasets within which individual patient predictions could be made for 90 day arterial thrombotic risk and 90 day major haemorrhage risk (see Figure 9-18 to Figure 9-23 below). For each trial a single imputed complete dataset was used to illustrate the general direction of model calibration in these data, whilst the values provided in Table 9-8 are pooled over the 10 imputed datasets.

Since *Trial 1* was used in the development of the two competing risk regressions it should be anticipated that the predicted risk matches the observed risk accurately. In this regard the calibration plots for *Trial 1* offer no real value in the assessment of model performance and are provided for completeness only.

For *Trials 2*, *5* and *7*, the prediction model for major haemorrhage consistently overestimated the risk of haemorrhage. In the case of *Trial 4* there was an indication that the major haemorrhage model underestimated those at low risk and overestimated those at high risk, whereas with *Trial 6*, the same model consistently underestimated the risk of haemorrhage.

The prediction model for arterial thrombosis was poorly calibrated within *Trials 2*, *6* and *7*. This model was well calibrated in *Trials 4* and *5*, although there is an indication that the model systematically overestimated the risk of arterial thrombosis.

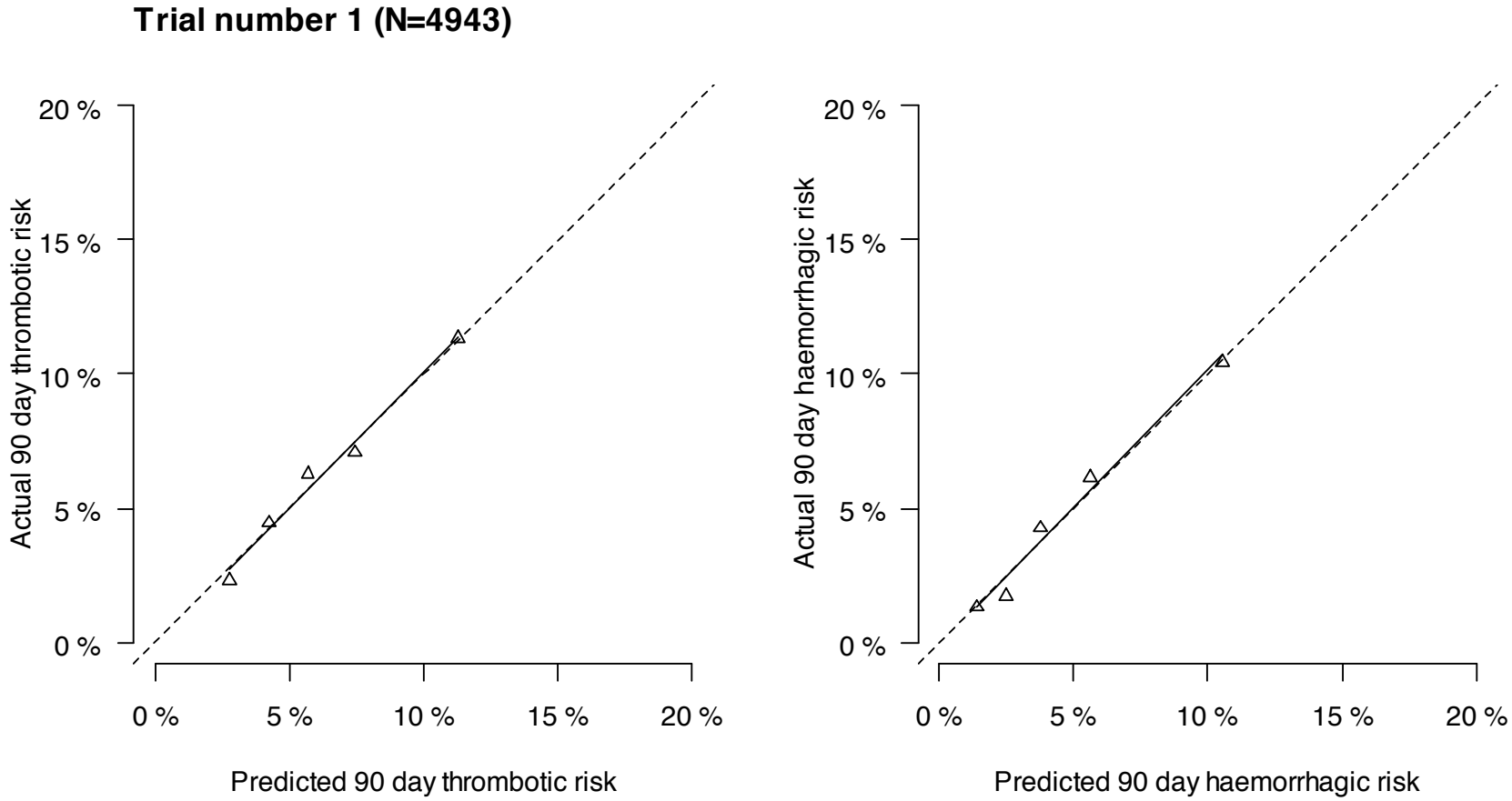


Figure 9-18 Calibration plot in development dataset (Trial 1) for arterial thrombosis and major haemorrhage.

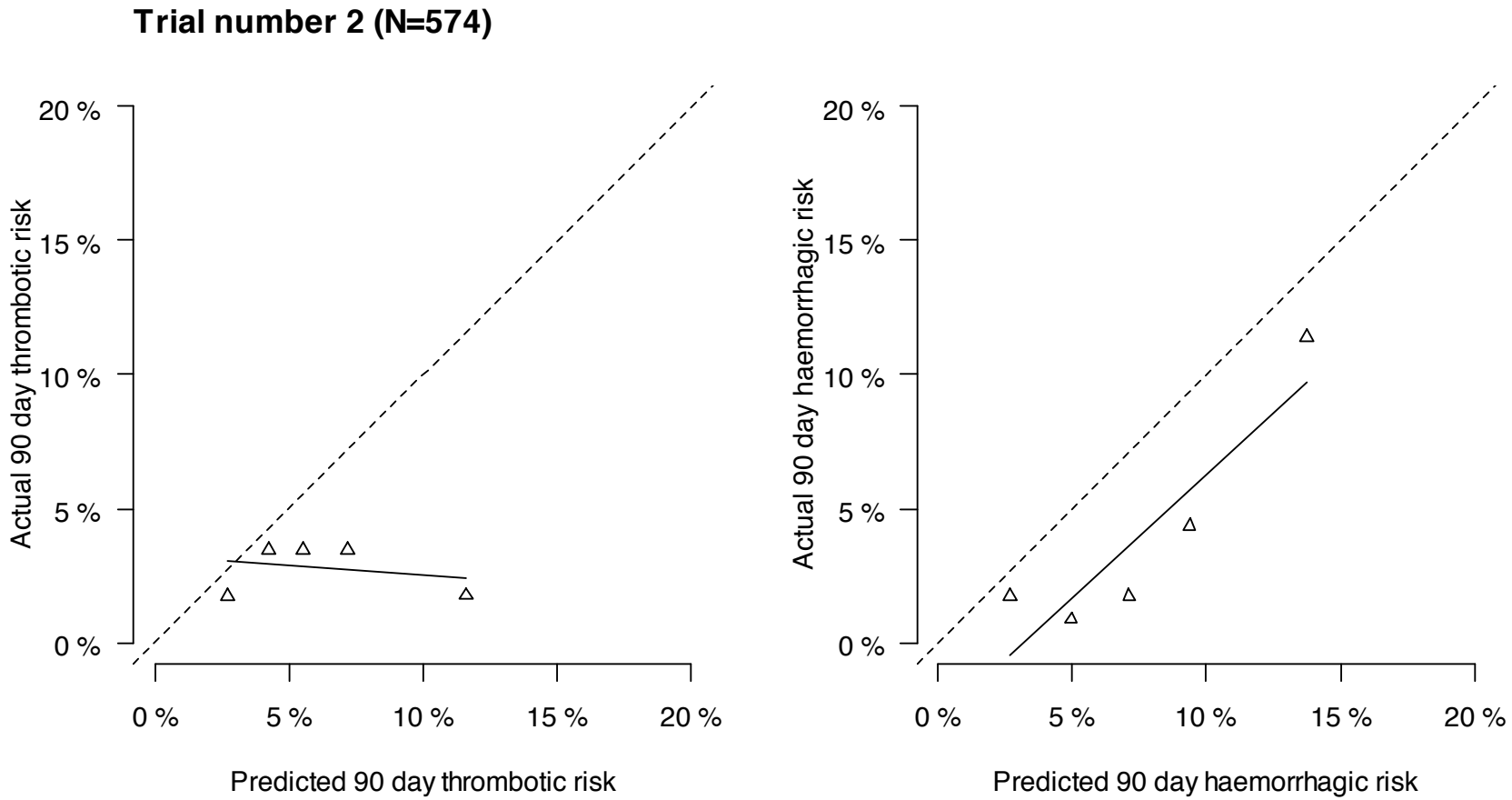


Figure 9-19 Calibration plot in external evaluation dataset (Trial 2) for arterial thrombosis and major haemorrhage.

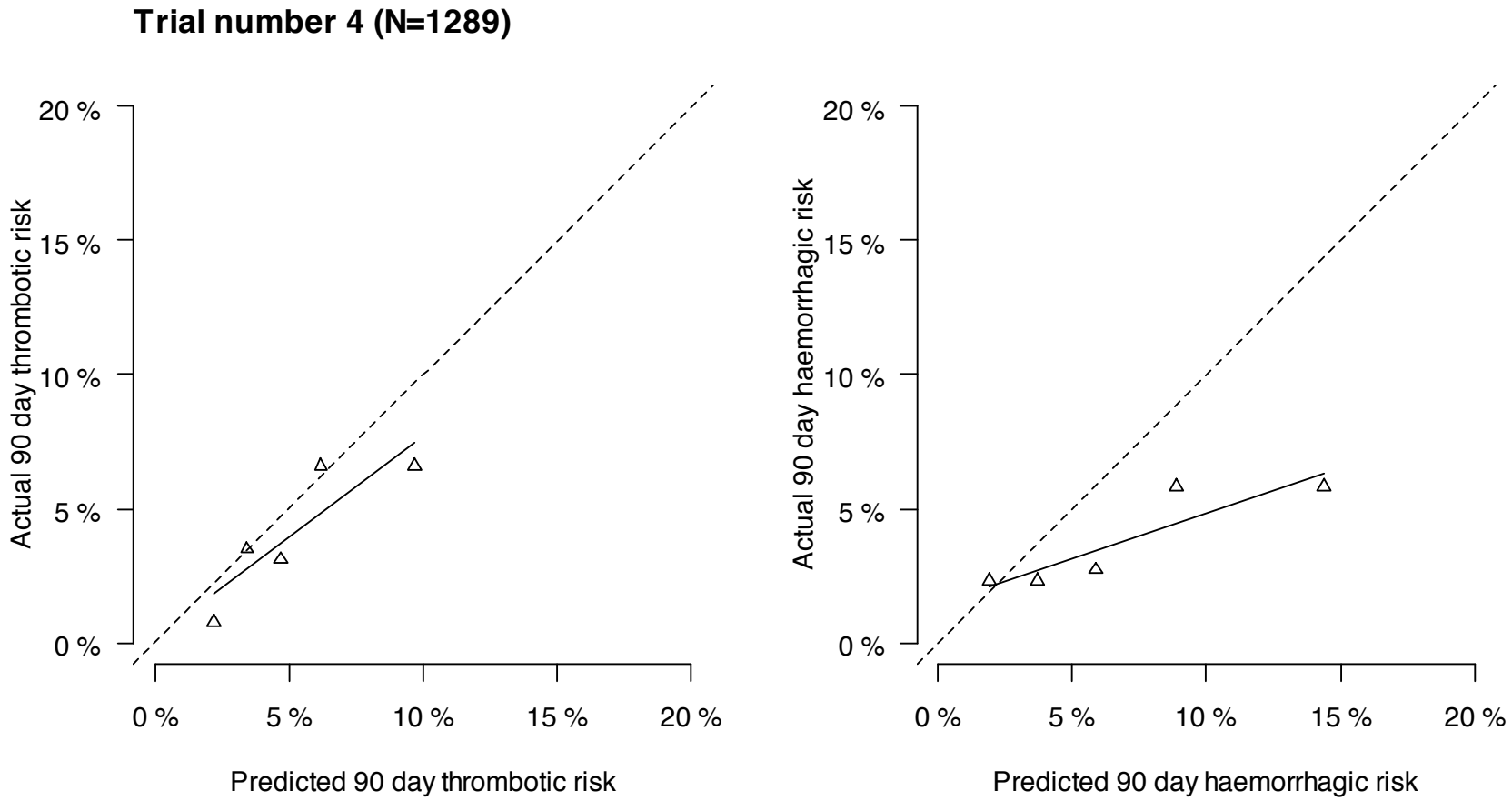


Figure 9-20 Calibration plot in external evaluation dataset (Trial 4) for arterial thrombosis and major haemorrhage.

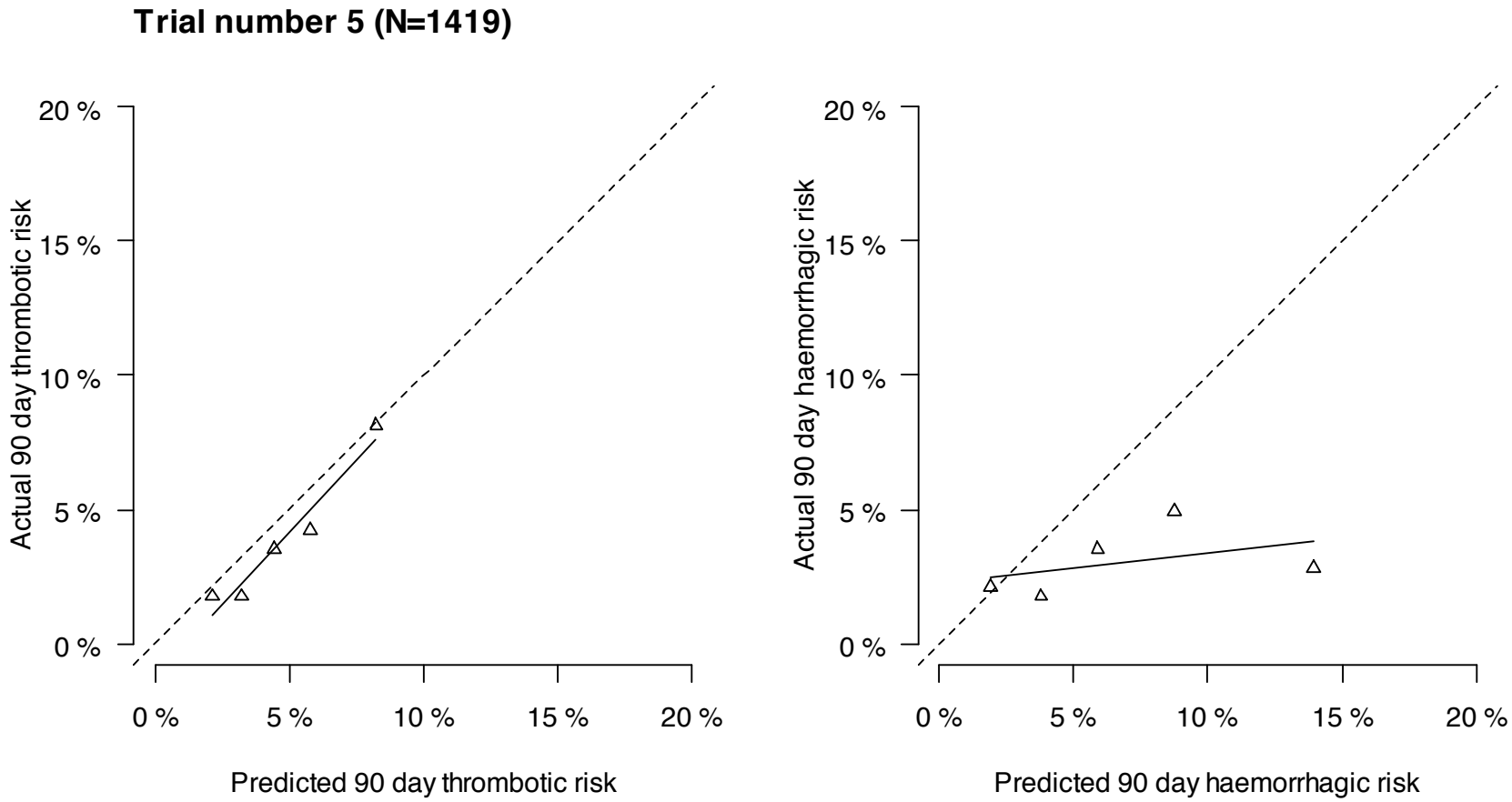


Figure 9-21 Calibration plot in external evaluation dataset (Trial 5) for arterial thrombosis and major haemorrhage.

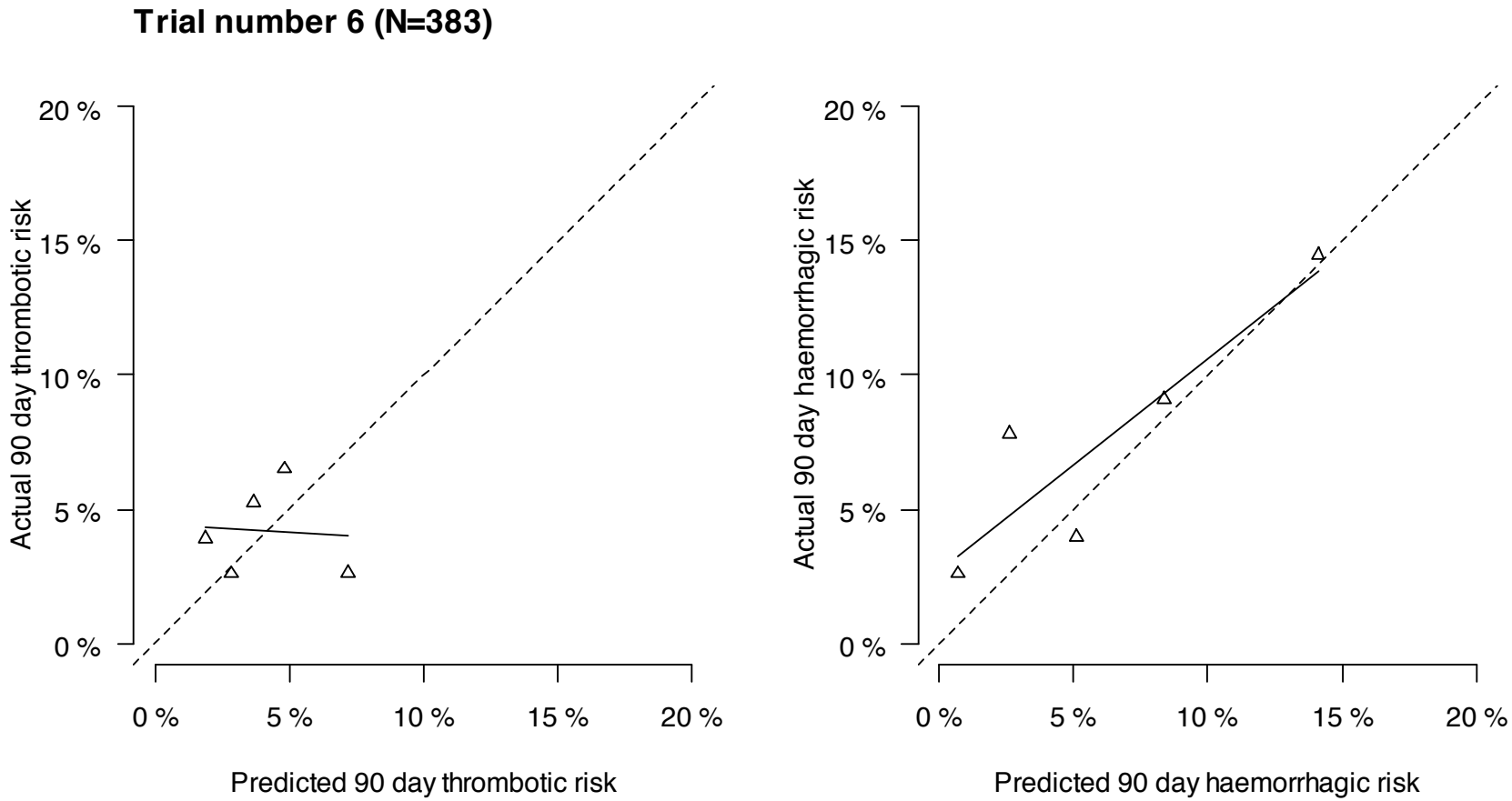


Figure 9-22 Calibration plot in external evaluation dataset (Trial 6) for arterial thrombosis and major haemorrhage.

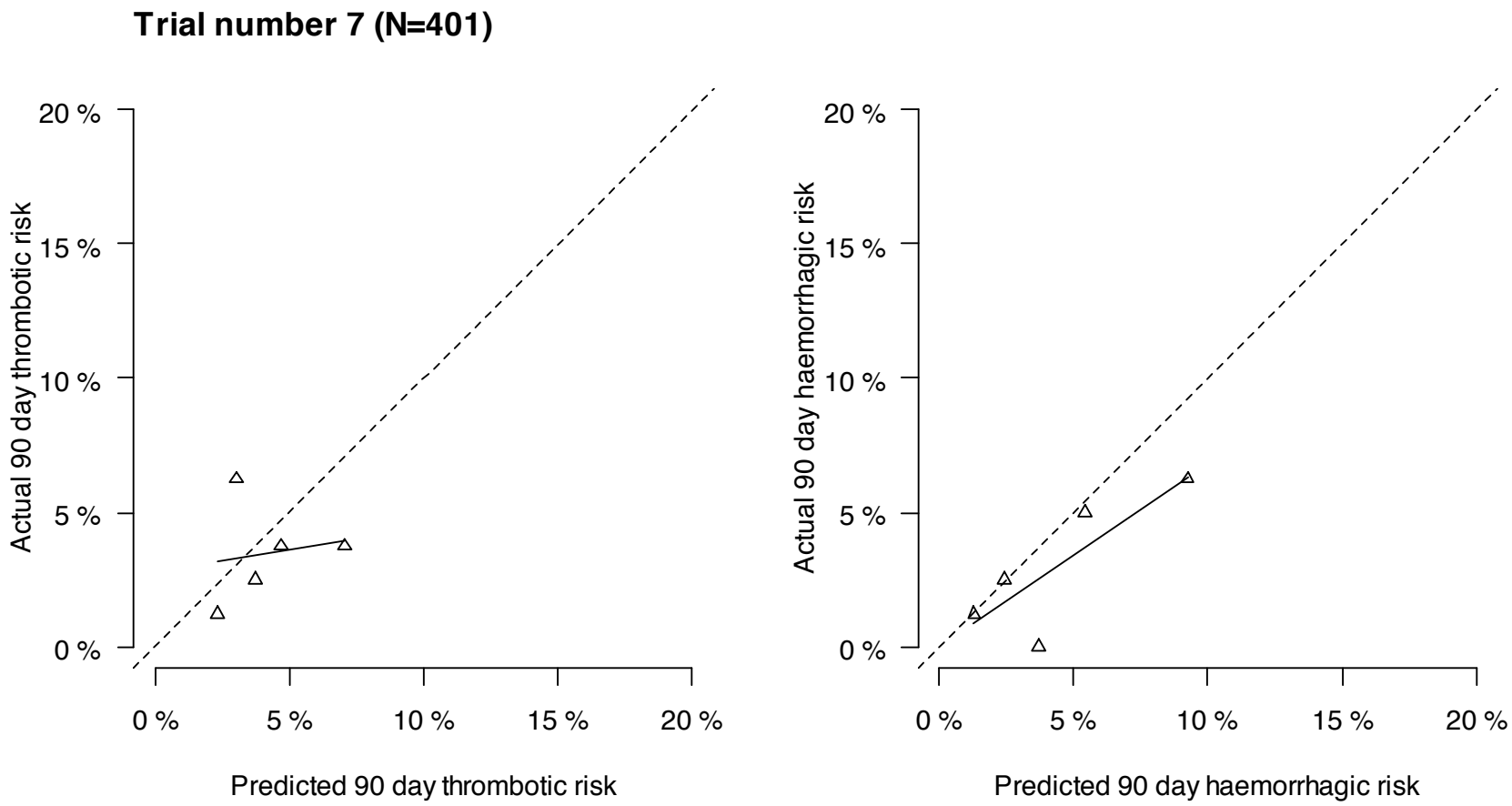


Figure 9-23 Calibration plot in external evaluation dataset (Trial 7) for arterial thrombosis and major haemorrhage.

9.10 Appendix D: Sensitivity analyses

There was a record of treatment with recombinant tissue plasminogen activator (rtPA) amongst those randomised in *Trials 1* and *3*. This comprised some 29% of the patients that made up the VISTA extract. It is not clear whether the other trials (i.e., *Trials 2*, and *4* through to *7*) suffer from a misreporting of a concomitant treatment with rtPA or else a true record of no rtPA. An assessment is now made as to what impact this makes to the Fine and Gray models. Additionally, an exclusion analysis is considered with all patients on rtPA removed from analysis. As *Trial 1* was one of the few trials with a record of concomitant rtPA treatment it is straightforward to make the required adjustment within the multivariable Fine and Gray models (see Table 9-7).

The effect of rtPA on *arterial thrombosis* was not significant with an adjusted sHR of 0.97 (95%CI 0.76 to 1.23) and a P-value of 0.8007. There were small differences in the magnitude of the other effects after including rtPA, though they were not substantial enough to change any conclusions previously drawn (see Figure 9-24). The effect of rtPA on *major haemorrhage* was significant at the 5% level with an adjusted sHR of 1.52 (95%CI 1.16 to 2.00) and a P-value of 0.0023. There was no qualitative difference in the magnitude of the other effects though the effect of increasing NIHSS was slightly less in the model which included rtPA. Additionally, the effect of prior use of any anticoagulants was slightly larger though remained non-significant (Figure 9-25).

Excluding those who received rtPA from the Fine and Gray models reduced the sample size from 4943 to 3028. This inflated the SEs which therefore widened the 95%CIs (see Figure 9-24 and Figure 9-25). For arterial thrombosis, the effect estimates for those characteristics related to medical history were further from the null effect whilst estimates for treatments were closer to the null. A similar inflation in the effect sizes was seen in the case of major haemorrhage. Diabetes was associated with a considerably larger increase in the hazard of haemorrhage as was NIHSS whilst current antiplatelets no longer increased the hazard of major haemorrhage and prior anticoagulants was associated with an increase in the risk.

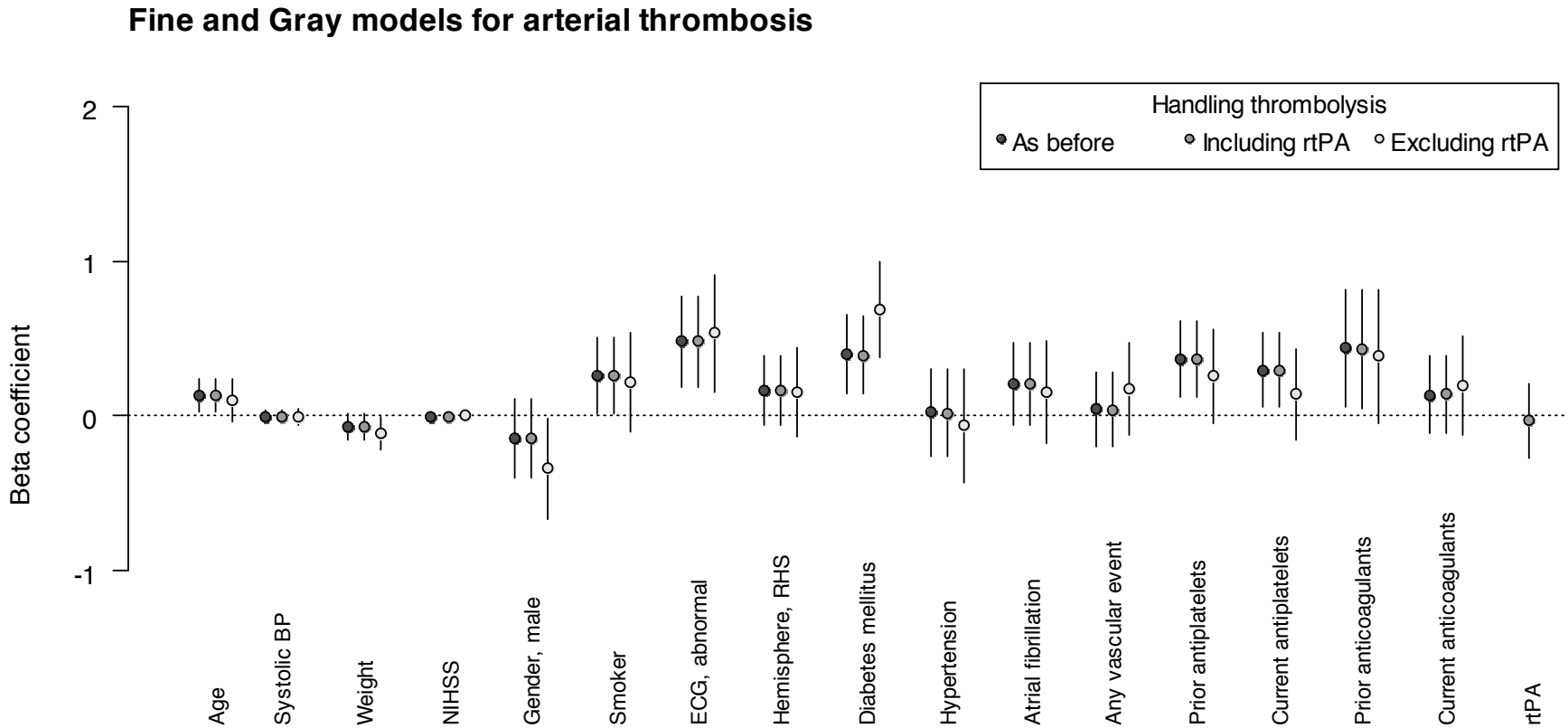


Figure 9-24 Forest plot of beta regression coefficients for a Fine and Gray arterial thrombosis model, handling thrombolysis in different ways.

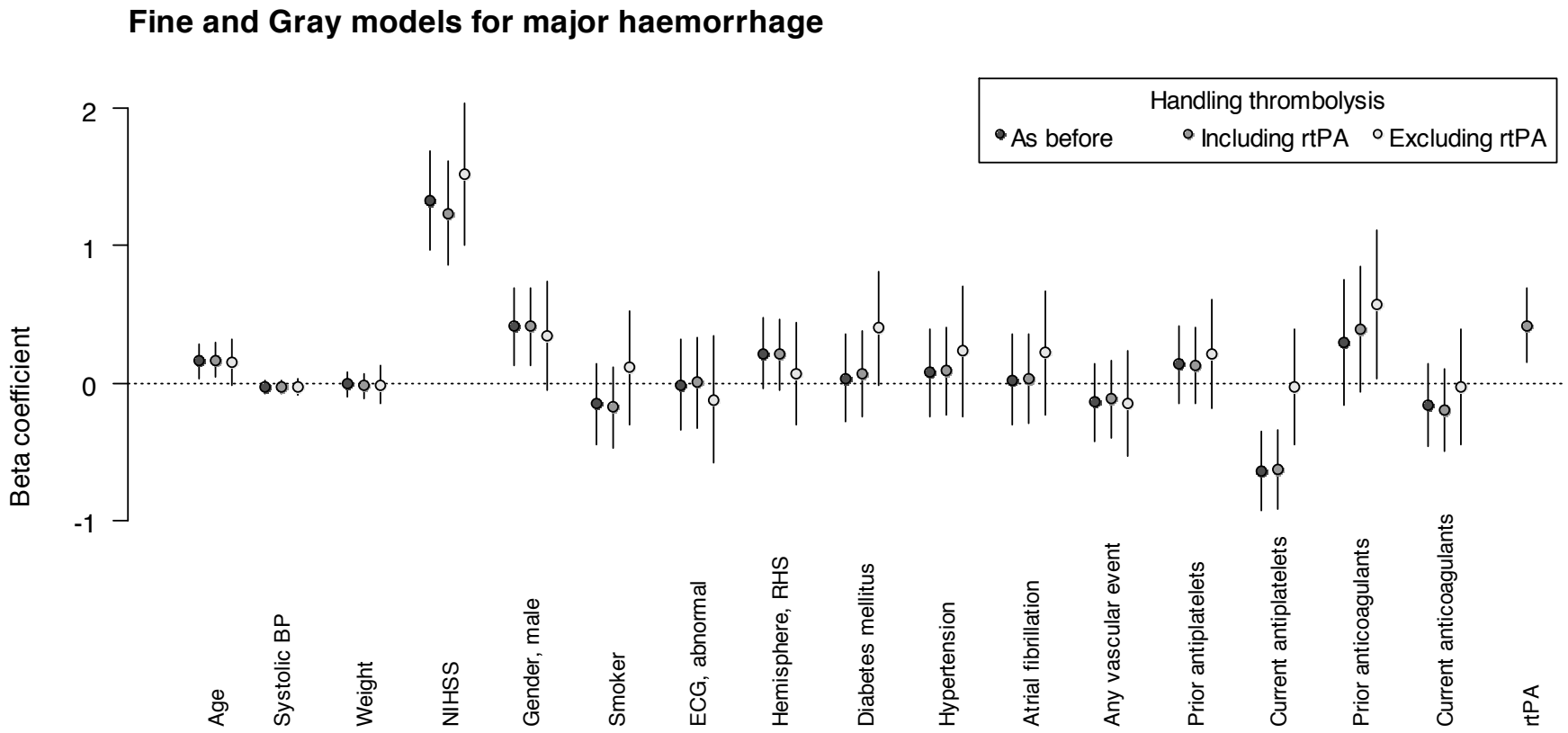


Figure 9-25 Forest plot of beta regression coefficients for a Fine and Gray for major haemorrhage model, handling thrombolysis in different ways.

Chapter 10: Discussion and Recommendations

Background and summary

This thesis has explored the use of clinical prediction models in stroke medicine. This work serves as an example as to how trial datasets can be better analysed to investigate more individualised estimates of a given treatment effect.

10.1 An overview of the thesis

This thesis has explored the possibility of a stratified framework for the treatment of ischaemic stroke patients based on their predicted risks of treatment harm and benefit. A critique of the methods of prediction was made contrasting formal with informal prediction to determine whether expert clinical gestalt differs from the predictions made by an objective clinical prediction model.

10.1.1 Methods of prediction in stroke

Recurrent events that follow a stroke in the short, medium and long term are important. Reliable predictions about prognosis would be useful in the management and treatment of patients. It was found that the clinical prediction models which are currently available for the prediction of recurrent stroke or MI (Chapter 3) suffered from the same methodological and reporting issues which persist in other medical conditions (Bouwmeester et al., 2012). Continuous predictors were frequently categorised thus discarding a considerable amount of predictive information. Authors favoured data-dependent variable selection techniques which were often applied to small samples of data with too few events to reliably estimate parameters. The quality of data was also poor with retrospective studies used and missing data poorly handled. A common theme was poor reporting: the consumers of research need to be informed about any decisions made by the researchers over the course of model development to allow the reader to critique any possible sources for bias. Despite

these weaknesses, a number of published evaluation studies suggested that some of these models achieved a moderate level of discriminative ability.

Expert opinion is the clear alternative to statistical prediction. A direct comparison of the two methods is a crucial – though rarely adopted – step in the process of model evaluation. Models identified in Chapter 3 were compared with clinicians' predictions in a prospective single centre observational study (Chapter 4). It was found that there was no real difference between them with regards discrimination suggesting that formal and informal methods of prediction not only discriminated between events and non-events moderately well but also did so similarly.

Models for the prediction of functional outcome (Chapter 5) were identified via a pre-existing systematic review – again these were compared to the predictions made by clinicians within the same observational cohort used in Chapter 4 (Veerbeek et al., 2011). Similar levels of ordinal discrimination were achieved suggesting that models and clinicians did as well as each other.

10.1.2 Application of clinical prediction models in RCTs

Exploring treatment interactions through one-at-a-time subgroup analyses is an inefficient method for investigating whether treatments benefit some patients but harm others. Multivariable risk prediction is a superior way of understanding patient heterogeneity and the consequent heterogeneity of the treatment effect (Kent et al., 2010, Pocock et al., 2014).

Three of the largest trials of aspirin in acute ischaemic stroke (i.e., IST-1, CAST and MAST-I) were re-analysed to explore the hypothesis that targeting treatment with aspirin to those patients at a high predicted risk of thrombosis (arterial or venous), a low predicted risk of haemorrhage, or a higher predicted risk of poor functional outcome would lead to overall benefit (Chapter 6). It was concluded that based on these datasets and the common simple clinical variables between them that there was no support for the stratified treatment of patients. It is likely that new prognostic variables would be required in order to ensure better discrimination of events.

A secondary analysis of the IST-3 dataset was presented in Chapter 7 exploring the stratified treatment of acute ischaemic stroke patients with rtPA according to their risk of treatment related harms (i.e., post rtPA SICH or poor functional outcome). Those at a high predicted risk of poor functional outcome were more likely to experience benefit with an indication that the beneficial relative effect of rtPA was greater with increasing risk. Paradoxically, those with a greater risk of post rtPA SICH also derived a larger treatment benefit. This is explained by the tendency for the two groups (i.e., high risk SICH and high risk poor outcome) to share similar attributes. Therefore, based on those existing clinical prediction models it appears that the stratification of treatment based on post rtPA SICH risk is not possible.

In Chapter 8 an exploratory analysis of the IST-3 dataset was presented. Here the impact of rtPA on mortality was investigated across an 18 month follow-up period whilst allowing for the possibility that treatment may interact with predicted prognosis at baseline and the delay in receiving treatment. These findings support the need for early administration of rtPA amongst those eligible and also support the treatment of those with a poor predicted prognosis at baseline. Although this analysis indicated that there is the potential for harm (increased risk of mortality) amongst those with a good prognosis (i.e., mild strokes) an emphasis should be placed on further randomised study of these patients before any firm conclusion can be drawn.

Finally, Chapter 9 explored the impact that secondary events (i.e., arterial thrombosis and major haemorrhage) have on mortality after an acute ischaemic stroke using data from the VISTA repository. A central aspect to this chapter was the illustration of a methodologically sophisticated approach to the handling of competing events. Again it was concluded that the predicted risks of thrombosis and haemorrhage based on simple clinical variables were too strongly correlated with the risk of other events like mortality or poor functional outcome.

10.2 Clinical Implications

The hypothesis that overall treatment benefit varies according to a patient's risk from treatment related harms is a biologically plausible one. Should there be any empirical evidence to support this hypothesis then a stratified approach would surely be advocated and adopted into clinical practice. In this thesis the use of pre-existing or newly developed clinical prediction models in randomised trial datasets failed to support such qualitative interactions under treatment with aspirin or treatment with rtPA in acute ischaemic stroke. However, if the current alternative is to base contraindications to treatment on single risk factors, each implicitly weighted by the treating clinician; then the lack of an interaction based on multivariable prediction techniques – a more formalised approach to combining the totality of clinical evidence – highlights concerns over the possible under-treatment of patients.

Thrombolysis with rtPA is one of the most effective treatments for acute ischaemic stroke; though the current licences held in Europe and the US are restricted. The Stroke Thrombolysis Trialists' Collaborative (STTC) group addressed these concerns in their IPD-MA (Emberson et al., 2014).They found that regardless of patient age, NIHSS score, and the heightened risk of SICH, rtPA significantly improved the chance of a good outcome when administered within a four and a half hour window from stroke onset. It is hoped that the work by the STTC group will now galvanise support for an extension in the current rtPA licences (Hill and Coutts, 2014). The STTC's conclusions corroborate the findings made in Chapter 7: the elderly and those that suffer more severe strokes are precisely those that stand to gain the largest benefit from rtPA.

Aspirin administered within 48 hours of acute ischaemic stroke onset is associated with a small but beneficial reduction in the risk of death or dependency. It was argued that the perceived risk of haemorrhage is best understood by considering the risk profile of a patient and that such a risk profile is multifaceted. Although the IPD-MA presented in Chapter 6 failed to identify those most likely to experience overall benefit or overall harm; the conclusion drawn is one of generalisability. This point was highlighted by Rothwell who regarded the lack of evidence for a subgroup

difference as an important finding stressing that this could indicate that treatment is more generalisable than previously thought (Rothwell, 2007b). The analysis in Chapter 6 supports the wide-spread use of aspirin immediately following an acute ischaemic stroke placing more emphasis on the benefits of early treatment with aspirin despite the risk of harm which could not be predicted using simple clinical variables. Aspirin remains an important treatment for improving the outcomes of acute ischaemic stroke patients especially amongst those who are currently ineligible for treatment with rtPA.

This thesis has drawn focus on the limitations of current prediction models, especially where simple clinical risk factors have been used. What role should prediction then play in stroke medicine? It will be argued here that the findings made in this thesis should be interpreted not as reason against using clinical prediction models in stroke but instead as a motivation to do better. For instance, it is evident that the prediction of poor functional outcome after an acute ischaemic stroke can be achieved using simple clinical risk factors producing models which perform exceptionally well in external evaluation. This was not the case though when predicting intermediate events, e.g., thromboses or haemorrhages. Here improvements in prediction may require novel predictors enabling better separation of the intermediate events from those with poor functional outcome. Such variables were not recorded reflecting the pragmatism adopted in the running of these large trials (recall that together IST-1 and CAST randomised around 40,000 patients). For example there was no record of: previous gastrointestinal ulceration; prior DVT; cancer; cerebral microbleeds; more advanced brain imaging findings; or physiological or genetic markers of aspirin metabolism. These variables may be available in newer trials of antiplatelet agents for the treatment of acute stroke. Similar points apply to the prediction of post rtPA SICH and those models adopted in Chapter 7. All of these used simple baseline clinical and imaging variables, however variables not measured, e.g., advanced imaging methods, genotyping or blood biomarkers related to the pathophysiology of post-rtPA SICH may better predict response to treatment.

Many prediction models in stroke have yet to find a role in the clinic. For instance, it was concluded in Chapter 5 that clinical prediction models make predictions of poor functional outcome that are at least as good as informal predictions made by doctors. Given this good performance it is plausible that any one of these models could be implemented in the screening of stroke patients on admission to the clinic. This has the potential to streamline the running of a stroke clinic, making more efficient use of the specialists' time as well as more focussed targeting of resources. Trained hospital staff (e.g., nurses or junior medics) could enter patient information into a simple application run on a computer tablet or a smart phone which acts as an interface for the mathematical equation pre-specified by a prediction model. This would utilise the ability of the prediction model in making reproducible and objective predictions. The following observations are now made. A dichotomisation of the predicted risk is required – one which is tuned to achieve high sensitivity or specificity at some threshold in a given dataset. The selection of this threshold will require long and careful consideration and should be closely and continuously monitored to establish the impact on patient outcomes. It is unlikely that the same threshold would be effective in different centres and as a consequence a centre-by-centre specification may be required. Periodic evaluations of the prediction model and regular re-calibration could be scheduled using within-centre data. This process could be automated with updates automatically compiled and reported to the clinical team using the `markdown` package in R calling upon a pre-prepared R script applied to a securely structured dataset ultimately generating an HTML document with imbedded results (Allaire et al., 2014). A dedicated and password protected webpage on an institution's intranet could be created precisely for reporting this information back to the clinical team. The Six Simple Variables model would be an ideal candidate for this as it is well established and has strong 'face-validity' (Counsell et al., 2002). Additionally, a previous study has already illustrated that the SSV model could serve a purpose in hospital audits: adjusting for case differences so as to give a fair comparison of mortality rates between multiple hospitals (Weir et al., 2001).

10.3 Methodological Implications

The work presented in this thesis adds to the growing number of published secondary analyses which take a fresh look at existing trial datasets by adopting an analysis strategy that recognises the heterogeneity present in any given patient population. A patient's risk-profile is individual and unique. Distributed across a population, it provides a continuous description of the rate at which bad outcomes are likely to occur in the absence of treatment. This understanding has received considerable interest over the last 10 years or so, particularly amongst vascular medicine groups.

10.3.1 Related work

For example, Pocock *et al*. illustrated the harmful impact that early ischaemic events (MI) and haemorrhagic complications have on the rate of mortality amongst those with Acute Coronary Syndrome (ACS) (Pocock et al., 2010). A risk-based decision strategy for selecting treatment was presented using new multivariable logistic regression models developed within the Acute Catheterization and Urgent Intervention Triage Strategy (ACUITY) trial. The ACUITY trial was a trial of patients with ACS randomised to receive either: heparin plus a glycoprotein IIb/IIIa inhibitor; bivalirudin plus a glycoprotein IIb/IIIa inhibitor, or bivalirudin alone. Pocock *et al*. found that the predictors for MI and haemorrhage were, for the most part, distinct, thus enabling individual predictions to be made which supported the selective treatment of patients, whereby a balance could be struck between the predicted chance of benefit and the predicted risk of harm.

Salisbury *et al*. have also described a risk-stratified framework for the treatment of Acute Coronary Syndrome (ACS) patients for the prevention of ischaemic complications but instead treated with clopidogrel or prasugrel (Salisbury et al., 2013). The authors re-analysed the Trial to Assess Improvement in Therapeutic Outcomes by Optimizing Platelet Inhibition With Prasugrel-Thrombolysis in Myocardial Infarction 38 (TRITON-TIMI 38). They obtained individual estimates of the risk of major/minor bleeding events and major ischaemia by developing new multivariable prediction models which adjusted for a number of pre-identified clinical risk factors and also allowed for treatment interactions. They found

considerable variation in the predicted response to treatment (harms and benefits) concluding that it was indeed possible to maximise the beneficial payoff from treatment whilst minimising the potential for harm.

Dorresteijn *et al*. re-analysed the Justification for the Use of Statins in Prevention (JUPTER) trial of 20mg rouvastatin versus placebo for the primary prevention of cardiovascular disease (i.e., MI, stroke etc.) (Dorresteijn et al., 2011b). They demonstrated that the use of either existing models or a newly developed model applied to the original trial data provided more net benefit in contrast to the two alternative treatment strategies, i.e., treat no-one; or treat everyone. In a separate, but similar analysis, Dorresteijn *et al*. re-analysed data from the Women's Health Study (WHS) trial for the primary reduction in the rate of vascular events (i.e., non-fatal MI or stroke, or a cardiovascular death) with 100mg aspirin (Dorresteijn et al., 2011a). Again, using either existing models or a newly developed model they illustrated the merits of selective treatment based on individualised risks. Note that in both of these publications Dorresteijn *et al*. used the net-benefit approach described by Vickers *et al*. which works by varying the threshold for deciding whether or not to treat (Dorresteijn et al., 2011a, Dorresteijn et al., 2011b, Vickers and Elkin, 2006). The following stances are then contrast with one another for the assessment of net benefit: (i) treat all; (ii) treat none; or (iii) treat some according to risk prediction.

A review by van der Leeuw *et al*. highlighted that one of the issues associated with weighting benefits and harms using individualised predicted risks is that frequently harmful effects will occur at a much lower prevalence requiring larger samples of patients to enable the accurate modelling of patient risk from harm (van der Leeuw et al., 2014). This is hardly surprising though, after all: a treatment with as great a risk of harm as the chance of benefit would be unlikely to reach approval from the FDA or a similar drug authority let alone garner enough interest to reach evaluation within a phase III trial (Note though that this entirely depends upon factors such as severity and chronicity of the disease). Additionally, van der Leeuw *et al*. point out that often the predicted risk of harm is intrinsically related to the predicted chance of benefit, i.e., those at a high predicted risk of treatment harm are the same that will derive the most benefit. This was indeed the case with many of the analyses explored in this

thesis, which as discussed earlier in this chapter suggests that the targeted treatment of stroke patients by their predicted risk is currently not a viable option. This is arguably the main barrier in succeeding with the risk-stratified approach in the treatment of acute stroke patients. Similar findings have been made elsewhere, for example, an IPD-MA which re-analysed five heparin RCTs in acute ischaemic stroke found that there was no evidence to suggest targeting the treatment of patients according to their predicted risks of treatment harm or benefit (Whiteley et al., 2013).

Since Kent *et al*. originally published their proposal in 2010 there has been a considerable amount of interest in this area with many researchers implementing the use of multivariable regression techniques to explore the heterogeneity of treatment effects in old datasets (Kent et al., 2010). More recently, an article by Pocock *et al*. again stressed the benefits of this approach calling on more trialists to include such methods in their publications (Pocock et al., 2014). This method has the potential to replace the inefficient one-at-a-time subgroup analysis approach; however, it is entirely dependent upon the availability of existing background information with regards to patient prognosis – though in most medical conditions some prior knowledge will be available. The work presented not only in this thesis but also across the various publications summarised above have provided extensive insight into the processes of harms and benefits associated with randomised treatments. It seems possible that such post-hoc analyses could be used to encourage trialists to design their next RCT to suitably test for a plausible treatment interaction with the predicted risk of harm. Of course to ensure suitable statistical power for the interaction test it is inevitable that the sample size must be inflated by some factor (Brookes et al., 2004). A direction for future research could explore the required increase in sample size when conditioning on the predictive accuracy of a given clinical prediction model. This could for instance be investigated using a simulation approach. If it were to be found that the factor of inflation was too great then this could be used to encourage more international collaboration and the sharing of existing datasets.

An ongoing question is how the results from such an analysis can most effectively be translated into the clinic and how patients may be best informed about the merits of

one medication over another (Pocock et al., 2014). Graphical presentation of risk is essential but given the inherent subjectivity involved in interpretation it is important that the most appropriate methods are used (Spiegelhalter, 2008, Spiegelhalter et al., 2011).

10.3.2 Simplicity versus complexity

A variety of clinical prediction models have been identified during this research. The methods applied in development ranged from those which used no statistical analysis (e.g., the SPAN-100 score derived using the unweighted sum of patient's age and NIHSS score arbitrarily categorised at 100 declaring any patients above that threshold as 'SPAN positive') to the development of new prediction models adopting more sophisticated methodology (e.g., the FG models developed in Chapter 9). This raises some interesting points regarding simplicity and complexity (see Chapter 2).

The large majority of pre-existing models evaluated in Chapter 7 were presented as point risk scores: a simplification of the regression equation achieved by finding a common denominator and rounding the estimated effects (Steyerberg, 2009). It might be anticipated that as a result of a loss in accuracy that is associated with the point risk score approach a regression equation would achieve better performance. In the IST-3 data the DRAGON score (an 11 point risk score) achieved an AUROCC of 0.78 whilst the König *et al*. model (a logistic regression model) achieved an AUROCC of 0.80. Therefore, despite the associated loss in accuracy, the two models performed just as well as one another at separating events from non-events. The calibration slopes of the two models were also good, suggesting that the weights adopted for each of the predictors were on average smaller in these data though still good (≈1). Calibration-in-the-large suggested that on average the DRAGON score overpredicted patients risk (intercept of 0.20) whilst the König model was a near perfect match on average (intercept of 0.03).

Here it is argued that there are considerable drawbacks associated with model developers not sharing the unique regression equation with original coefficients. First, this approach can preclude others from making accurate assessments of model performance in new data. Secondly, it stands in the way of model updating (or re-

calibration); an approach which brings an existing model in-line with new data (Steyerberg, 2009). Model updating can range from subtle changes which correct the systematic component of the model (i.e., the intercept or baseline hazard which anchors the linear predictor) to completely re-estimating the weights attributed to each of the included predictors. In any case, model updating makes better use of the existing scientific knowledge (i.e., the prediction model already available) and the new knowledge acquired (i.e., the new data). This should be favoured over simply developing a new model whenever an existing model does not yield a good fit (Moons et al., 2012b).

If authors cannot be persuaded to present this information in the body of their publication then they should be encouraged to provide it as an online supplement. By sharing these data authors are allowing more efficient use of their work. Explicitly reporting the baseline hazard in the case of proportional hazards models derived from time-to-event data (e.g., the Cox PHM) is slightly more complicated and requires more information to be reported. This is discussed in detail by Royston and Altman, who note that whilst a complete expression for the baseline hazard cannot be provided, a smoothed estimate can be obtained using spline functions or Fractional Polynomials (Royston and Altman, 2013). In this form the baseline hazard can then be explicitly written down. One drawback of this is that expert statistical input is required, but it might be argued that this should nevertheless be essential, and mandating such practice might improve the quality of the published literature.

In any case, with the volume of published prediction models increasing and methodologists highlighting new methods for aggregating multiple models, the developers of new prediction models must be as explicit as possible when reporting their models to enable better assessment of the possible biases involved as well as enabling more efficient use of the model itself (Debray et al., 2014, Collins, 2011).

10.4 Limitations

A general comment is made here which holds true of all of the analyses presented in this thesis. It is invariably impossible that a common consensus in science is reached and it is therefore very likely that some will disagree with approaches taken both with the analysis and with the handling of these datasets. The stance taken throughout this thesis has been one of inclusivity. The analyses that were favoured and therefore presented as *primary* were based on the majority of patients. It is inevitable then that the opinions of the reader and the author may, on occasion, be split. However, sensitivity analyses have been utilised throughout in an attempt to shed some light on areas of disagreement. It is the author's opinion that it is preferable to let the data 'speak' and allow the reader to make up his or her own mind as to whether they agree or disagree with the inferences that have been drawn.

Some issues regarding the quality of the analysed datasets are now considered.

Across its life span – pre and post publication – the IST-3 trial has attracted a considerable amount of interest (Lyden, 2012). Following the publication of the primary results in 2012 various critics voiced their concerns about the findings, citing in particular issues with design and analysis (Hoffman and Cooper, 2012, Fatovich, 2012, Fatovich et al., 2012). There are indeed a number of acknowledged limitations associated with the IST-3 (Lyden, 2012). Firstly (apart from an initially blinded pilot phase) it was an un-blinded trial, i.e.: the patient; any proxies; and the treating staff knew whether or not rtPA had been administered. The investigators attempted to mitigate this problem by blinding the assessment of six month functional outcome from treatment allocation. However, this may have had an influence on the patient's (or proxy's) responses to the interviewer. Secondly, the use of the 'uncertainty principle' meant that no distinct inclusion or exclusion criteria were enforced. If the treating physician was certain that the patient *should* receive rtPA then they should be treated accordingly and *not randomised*, likewise if the physician thought that the patient *should not* receive rtPA then, again, the patient should be treated accordingly and *not randomised*. Only if the physician was suitably uncertain then the patient *should be randomised*. Finally, overall the IST-3 was a neutral trial with no

statistically significant treatment difference in the primary outcome (OHS 0-2 vs. 3-6). However, the key secondary outcome, which adopted the more statistically efficient POLR approach in the analysis of ordinal outcomes, did demonstrate a favourable shift in disability scores at both 6 and 18 months.

It is wrong to simply discard the IST-3 out of hand because of its limitations; especially when one considers the efforts made by the trialists and the patients. The harshest critic could not deny that even at its worst IST-3 is an exceptional survey of similar patients in receipt of rtPA or standard care – there may be some bias associated with the effect of rtPA but the data are still of great value. At its best though the IST-3 is a fully randomised clinical trial offering fascinating insights into the effects of rtPA in patient groups previously untested. The IST-3 has undoubtedly made considerable additions to the evidence base for thrombolytics in acute ischaemic stroke.

One possible concern with the aspirin IPD-MA presented in Chapter 6 is the *age* of the trials with some of the patients recruited as long ago as January 1991. However, this raises an arguably more general question which should be considered first: when do data become no-longer relevant (e.g., through a secular shift in clinical practice) and become historical artefacts? A formal investigation of this might adopt a meta-regression analysis assessing whether a trend could be attributed to the date of publication (or perhaps date of recruitment start/end). Specifically though in the case of the aspirin trials, it is noted from the most recent Cochrane review of antiplatelets in acute ischaemic stroke that only eight trials were pooled with no new trials added since the previous review was undertaken in 2008 (Sandercock et al., 2014, Sandercock et al., 2008). In fact in the 2014 review the authors state that 98% of the total available data are made up by just two of the included trials (notably IST-1 and CAST). In short, even if there was concern that the data are somehow no longer relevant, it remains the case that there is an insufficient amount of trial-level data to properly test this hypothesis.

The issue of 'trial age' also placed an inherent restriction on the availability of predictors. For instance, stroke severity scales (e.g., the NIHSS, the SSS etc.) are far

more common in modern trials, however, in IST-1, CAST and MAST-I only single binary deficit measures were available. This limits comparisons to other more modern stroke trials; for the purpose of the analysis presented though it is likely that the same conclusions would have been drawn had for example the NIHSS been recorded. In any case, measures of stroke deficits are strongly related with one another and the single deficits used in these aspirin trials likely captured most of the prognostic information (Goldie et al., 2014).

Additionally, aspects of trial design differed between the aspirin datasets (e.g.: duration of early follow-up and consequently the assessment of endpoints; the dosage of aspirin administered and the dosage time etc.). In particular, these were large pragmatic trials where the primary endpoint was 'poor functional outcome' at six months. It is very likely that there is an under ascertainment of early secondary events which is therefore a potential source of bias. All of these issues may have diluted the effect of aspirin. However, considering each trial on its own, none were adequately powered to detect a treatment by benefit/harm interaction if it exists. The pooling of these datasets was therefore necessary increasing the sample size and the statistical power. Indeed this handling of the datasets matched the Cochrane antiplatelet review.

More generally, model development and application is entirely dependent upon the availability of patient information, i.e., the model inputs. It was often the case in this thesis that either existing models could not be assessed in the available data (i.e., Chapters 4 and 5) or else variable selection in model development was limited by the record of clinical risk factors (i.e., Chapters 6, 7, 8 and 9).

Finally, one of the main limitations of the analyses presented in Chapters 6, 7, 8 and 9 is that they were all post-hoc and unplanned. None of these trials were designed to test the hypotheses explored in this thesis and therefore caution is advised when interpreting these findings. It is noted however that this is true of any similar such analysis.

10.5 Future work

This thesis provides a strong argument for exploring the formal impact that any one of the prediction models presented in Chapter 5 would have in clinical practice. A so called 'model impact study' could be designed to assess whether the explicit use of a prediction model for the risk of poor functional outcome after stroke yields benefits for patients (e.g., improved recovery or quality of life outcomes) and/or savings for the health care providers (e.g., better use of resources be that economic or time) (Moons et al., 2012a). One candidate for this would be the SSV model, a few reasons are given: first, the external performance of the SSV model has already been illustrated; second, there is evidence to suggest that it discriminates just as well as a clinician (i.e., from Chapter 5 and the work by Counsell *et al*.); and finally it has good face validity (Counsell et al., 2004).

It was consistently demonstrated in Chapters 6, 7 and 9 that the predicted risk of intermediate events is strongly related to longer term outcomes like patient disability or mortality. The identification of new risk factors through etiological research to uniquely characterise those at high risk from haemorrhage but low risk from thrombosis must be identified before the stratified treatment of stroke patients can progress.

In Chapter 9 competing events were explored. A more sophisticated methodological approach using multi-state modelling could help elucidate whether risk factors and the qualitative or quantitative strength of the predictor-outcome associates differ in distinguishing those who suffer either a haemorrhagic event or a thrombotic event and then either survive or die in follow-up. The incidence of these intermediate events is small placing some restrictions on the availability of sufficient statistical power to obtain reliable parameter estimates especially under a multi-state framework. Therefore, access to data obtained under a large stroke registry or perhaps through multiple stroke centres, is required. Recent collaborative efforts of the like conducted by the Farr Institute of Health Informatics Research could allow such research utilising its extensive data linkage of health records from around the UK.

References

ADAMS, H. P., DEL ZOPPO, G., ALBERTS, M. J., BHATT, D. L., BRASS, L., FURLAN, A., GRUBB, R. L., HIGASHIDA, R. T., JAUCH, E. C., KIDWELL, C., LYDEN, P. D., MORGENSTERN, L. B., QURESHI, A. I., ROSENWASSER, R. H., SCOTT, P. A. & WIJDICKS, E. F. M. 2007. Guidelines for the Early Management of Adults With Ischemic Stroke: A Guideline From the American Heart Association/ American Stroke Association Stroke Council, Clinical Cardiology Council, Cardiovascular Radiology and Intervention Council, and the Atherosclerotic Peripheral Vascular Disease and Quality of Care Outcomes in Research Interdisciplinary Working Groups: The American Academy of Neurology affirms the value of this guideline as an educational tool for neurologists. *Stroke,* 38**,** 1655-1711.

AGRESTI, A. 2002. *Categorical data analysis*, John Wiley & Sons.

AHMED, I., DEBRAY, T., MOONS, K. & RILEY, R. 2014. Developing and validating risk prediction models in an individual participant data meta-analysis. *BMC Medical Research Methodology,* 14**,** 3.

ALBERS, G. W., CAPLAN, L. R., EASTON, J. D., FAYAD, P. B., MOHR, J. P., SAVER, J. L. & SHERMAN, D. G. 2002. Transient Ischemic Attack — Proposal for a New Definition. *New England Journal of Medicine,* 347**,** 1713-1716.

ALI, M., BATH, P. M. W., CURRAM, J., DAVIS, S. M., DIENER, H.-C., DONNAN, G. A., FISHER, M., GREGSON, B. A., GROTTA, J., HACKE, W., HENNERICI, M. G., HOMMEL, M., KASTE, M., MARLER, J. R., SACCO, R. L., TEAL, P., WAHLGREN, N.-G., WARACH, S., WEIR, C. J. & LEES, K. R. 2007. The Virtual International Stroke Trials Archive. *Stroke,* 38**,** 1905-1910.

ALLAIRE, J., HORNER, J., MARTI, V. & PORTE, N. 2014. markdown: Markdown rendering for R.

ALTMAN, D. G. 2009. Prognostic models: a methodological framework and review of models for breast cancer. *Cancer Investigation,* 27**,** 235-243.

ALTMAN, D. G. & DE STAVOLA, B. L. 1994. Practical problems in fitting a proportional hazards model to data with udated measurements of the covariates. *Statistics in Medicine,* 13**,** 301-341.

ALTMAN, D. G. & ROYSTON, P. 2006. The cost of dichotomising continuous variables. *Bmj,* 332**,** 1080.

ALTMAN, D. G., VERGOUWE, Y., ROYSTON, P. & MOONS, K. G. M. 2009. Prognosis and prognostic research: validating a prognostic model. *Bmj,* 338.

ÁLVAREZ-SABÍN, J., MAISTERRA, O., SANTAMARINA, E. & KASE, C. S. 2013. Factors influencing haemorrhagic transformation in ischaemic stroke. *The Lancet Neurology,* 12**,** 689-705.

ALVAREZ-SABIN, J., QUINTANA, M., RODRIGUEZ, M., ARBOIX, A., RAMIREZ, J. M. & FUENTES, B. 2008. Validation of the Essen risk scale and its adaptation to the Spanish population. Modified Essen risk scale. *Neurologia,* 23**,** 209-14.

ANTITHROMBOTIC-TRIALISTS'-COLLABORATION 2002. Collaborative meta-analysis of randomised trials of antiplatelet therapy for prevention of death, myocardial infarction, and stroke in high risk patients. *Bmj,* 324**,** 71-86.

APPELROS, P., NYDEVIK, I. & VIITANEN, M. 2003. Poor Outcome After First-Ever Stroke: Predictors for Death, Dependency, and Recurrent Stroke Within the First Year. *Stroke,* 34**,** 122-126.

ARIESEN, M. J., CLAUS, S. P., RINKEL, G. J. E. & ALGRA, A. 2003. Risk Factors for Intracerebral Hemorrhage in the General Population: A Systematic Review. *Stroke,* 34**,** 2060-2065.

ARMSTRONG, J. R. & MOSHER, B. D. 2011. Aspiration Pneumonia After Stroke: Intervention and Prevention. *The Neurohospitalist,* 1**,** 85-93.

AY, H., GUNGOR, L., ARSAVA, E. M., ROSAND, J., VANGEL, M., BENNER, T., SCHWAMM, L. H., FURIE, K. L., KOROSHETZ, W. J. & SORENSEN, A. G. 2010. A score to predict early risk of recurrence after ischemic stroke. *Neurology,* 74**,** 128-135.

AZARPAZHOOH, M. R., NICOL, M. B., DONNAN, G. A., DEWEY, H. M., STURM, J. W., MACDONELL, R. A. L., PEARCE, D. C. & THRIFT, A. G. 2008. Patterns of stroke recurrence according to subtype of first stroke event: the North East Melbourne Stroke Incidence Study (NEMESIS). *International Journal of Stroke,* 3**,** 158-164.

AZUR, M. J., STUART, E. A., FRANGAKIS, C. & LEAF, P. J. 2011. Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research,* 20**,** 40-49.

BABYAK, M. A. 2004. What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models. *Psychosomatic Medicine,* 66**,** 411-421.

BACK, A. L. & ARNOLD, R. M. 2006. Discussing Prognosis: "How Much Do You Want to Know?" Talking to Patients Who Are Prepared for Explicit Information. *Journal of Clinical Oncology,* 24**,** 4209-4213.

BAMFORD, J., SANDERCOCK, P., DENNIS, M., BURN, J. & WARLOW, C. 1990. A prospective study of acute cerebrovascular disease in the community:

the Oxfordshire Community Stroke Project--1981-86. 2. Incidence, case fatality rates and overall outcome at one year of cerebral infarction, primary intracerebral and subarachnoid haemorrhage. *Journal of Neurology, Neurosurgery & Psychiatry,* 53**,** 16-22.

BAMFORD, J. M., SANDERCOCK, P. A., WARLOW, C. P. & SLATTERY, J. 1989. Interobserver agreement for the assessment of handicap in stroke patients. *Stroke,* 20**,** 828.

BEDNAR, M. M. & GROSS, C. E. 1999. Antiplatelet Therapy in Acute Cerebral Ischemia. *Stroke,* 30**,** 887-893.

BENDER, R. & GROUVEN, U. 1997. Ordinal logistic regression in medical research. *Journal of the Royal College of Physicians of London,* 31**,** 546-551.

BIESHEUVEL, C. J., VERGOUWE, Y., STEYERBERG, E. W., GROBBEE, D. E. & MOONS, K. G. M. 2008. Polytomous logistic regression analysis could be applied more often in diagnostic research. *Journal of Clinical Epidemiology,* 61**,** 125-134.

BORENSTEIN, M., HEDGES, L. V., HIGGINS, J. P. T. & ROTHSTEIN, H. R. 2010. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods,* 1**,** 97-111.

BORNSTEIN, N., SILVESTRELLI, G., CASO, V. & PARNETTI, L. 2006. Arterial Hypertension and Stroke Prevention: An Update. *Clinical and Experimental Hypertension,* 28**,** 317-326.

BOUWMEESTER, W., ZUITHOFF, N. P. A., MALLETT, S., GEERLINGS, M. I., VERGOUWE, Y., STEYERBERG, E. W., ALTMAN, D. G. & MOONS, K. G. M. 2012. Reporting and Methods in Clinical Prediction Research: A Systematic Review. *PLoS Med,* 9**,** e1001221.

BOX, G. E. & DRAPER, N. R. 1987. *Empirical model-building and response surfaces*, John Wiley & Sons.

BROOKES, S. T., WHITELY, E., EGGER, M., SMITH, G. D., MULHERAN, P. A. & PETERS, T. J. 2004. Subgroup analyses in randomized trials: risks of subgroup-specific analyses;: power and sample size for the interaction test. *Journal of Clinical Epidemiology,* 57**,** 229-236.

BROWN, H. & PRESCOTT, R. 2006. *Applied mixed models in medicine*, John Wiley & Sons.

BURN, J., DENNIS, M., BAMFORD, J., SANDERCOCK, P., WADE, D. & WARLOW, C. 1994. Long-term risk of recurrent stroke after a first-ever stroke. The Oxfordshire Community Stroke Project [published erratum appears in Stroke 1994 Sep;25(9):1887]. *Stroke,* 25**,** 333-337.

CALIFF, R. M., WOODLIEF, L. H., HARRELL JR, F. E., LEE, K. L., WHITE, H. D., GUERCI, A., BARBASH, G. I., SIMES, R. J., WEAVER, W. D. D., SIMOONS, M. L. & TOPOL, E. J. 1997. Selection of thrombolytic therapy for individual patients: Development of a clinical model. *American Heart Journal,* 133**,** 630-639.

CHANDRATHEVA, A., GERAGHTY, O. C. & ROTHWELL, P. M. 2011. Poor performance of current prognostic scores for early risk of recurrence after minor stroke. *Stroke,* 42**,** 632-637.

CHEN, Z., SANDERCOCK, P., PAN, H., COUNSELL, C., COLLINS, R., LIU, L., XIE, J., WARLOW, C., PETO, R., CAST, O. B. O. T. & GROUPS, I. C. 2000. Indications for Early Aspirin Use in Acute Ischemic Stroke: A Combined Analysis of 40 000 Randomized Patients From the Chinese Acute Stroke Trial and the International Stroke Trial. *Stroke,* 31**,** 1240-1249.

CHRISTAKIS, N. A. & LAMONT, E. B. 2000. Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort study. *Bmj,* 320**,** 469-473.

CHRISTIE, M., CLIFFE, A., DAWID, P. & SENN, S. 2011. *Simplicity, Complexity and Modelling*, John Wiley & Sons, Ltd.

CLARK, T. G., ALTMAN, D. G. & STAVOLA, B. L. D. 2002. Quantification of the completeness of follow-up. *The Lancet,* 359**,** 1309-1310.

COHEN, J. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin,* 70**,** 213.

COLE, S. R., CHU, H. & GREENLAND, S. 2014. Maximum Likelihood, Profile Likelihood, and Penalized Likelihood: A Primer. *American Journal of Epidemiology,* 179**,** 252-260.

COLLEN, D. & LIJNEN, H. R. 2009. The Tissue-Type Plasminogen Activator Story. *Arteriosclerosis, Thrombosis, and Vascular Biology,* 29**,** 1151-1155.

COLLETT, D. 2003. *Modelling survival data in medical research*, CRC press.

COLLINS, G. 2011. Opening up multivariable prediction models: Consensus-based guidelines for transparent reporting. *BMJ* [Online].

COLLINS, G., DE GROOT, J., DUTTON, S., OMAR, O., SHANYINDE, M., TAJAR, A., VOYSEY, M., WHARTON, R., YU, L.-M., MOONS, K. & ALTMAN, D. 2014. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Medical Research Methodology,* 14**,** 40.

COLLINS, G. & LE MANACH, Y. 2013. Multivariable Risk Prediction Models: It's All about the Performance. *Anesthesiology,* 118**,** 1252-1253 10.1097/ALN.0b013e31828e13e9.

COLLINS, G. S., MALLETT, S., OMAR, O. & YU, L.-M. 2011. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Medicine*.

COOK, N. R. 2007. Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction. *Circulation,* 115**,** 928-935.

COUNSELL, C. & DENNIS, M. 2001. *Systematic review of prognostic models in patients with acute stroke*, Cerebrovascular Diseases. 12 (3) (pp 159-170), 2001. Date of Publication: 2001.

COUNSELL, C., DENNIS, M. & MCDOWALL, M. 2004. Predicting functional outcome in acute stroke: comparison of a simple six variable model with other predictive systems and informal clinical prediction. *Journal of Neurology, Neurosurgery & Psychiatry,* 75**,** 401-405.

COUNSELL, C., DENNIS, M., MCDOWALL, M. & WARLOW, C. 2002. Predicting Outcome After Acute and Subacute Stroke: Development and Validation of New Prognostic Models. *Stroke,* 33**,** 1041-1047.

COX, D. R. 1972. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological),* 34**,** 187-220.

COX, D. R. 1975. Partial likelihood. *Biometrika,* 62**,** 269-276.

COX, D. R. 1990. Role of Models in Statistical Analysis. *Statistical Science,* 5**,** 169-174.

CRAWFORD, F., ANANDAN, C., CHAPPELL, F., MURRAY, G., PRICE, J., SHEIKH, A., SIMPSON, C., MAXWELL, M., STANSBY, G., YOUNG, M., ABBOTT, C., BOULTON, A., BOYKO, E., KASTENBAUER, T., LEESE, G., MONAMI, M., MONTEIRO-SOARES, M., RITH-NAJARIAN, S., VEVES, A., COATES, N., JEFFCOATE, W., LEECH, N., FAHEY, T. & TIERNEY, J. 2013. Protocol for a systematic review and individual patient data meta-analysis of prognostic factors of foot ulceration in people with diabetes: the international research collaboration for the prediction of diabetic foot ulcerations (PODUS). *BMC Medical Research Methodology,* 13**,** 22.

CROWSON, C. S., ATKINSON, E. J. & THERNEAU, T. M. 2013. Assessing calibration of prognostic risk scores. *Statistical Methods in Medical Research*.

CUCCHIARA, B., KASNER, S., TANNE, D., LEVINE, S., DEMCHUK, A., MESSE, S., SANSING, L., LEES, K., LYDEN, P. & FOR THE, S. I. 2011.

Validation assessment of risk scores to predict postthrombolysis intracerebral haemorrhage. *International Journal of Stroke,* 6**,** 109-111.

DAWID, P. & SENN, S. 2011. Statistical Model Selection. *Simplicity, Complexity and Modelling*. John Wiley & Sons, Ltd.

DEBRAY, T. P. A., KOFFIJBERG, H., NIEBOER, D., VERGOUWE, Y., STEYERBERG, E. W. & MOONS, K. G. M. 2014. Meta-analysis and aggregation of multiple published prediction models. *Statistics in Medicine,* 33**,** 2341-2362.

DEBRAY, T. P. A., MOONS, K. G. M., AHMED, I., KOFFIJBERG, H. & RILEY, R. D. 2013. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Statistics in Medicine***,** n/a-n/a.

DEEKS, J. J., ALTMAN, D. G. & BRADBURN, M. J. 2008. Statistical Methods for Examining Heterogeneity and Combining Results from Several Studies in Meta-Analysis. *Systematic Reviews in Health Care*. BMJ Publishing Group.

DEHING-OBERIJE, C., DE RUYSSCHER, D., PETIT, S., VAN MEERBEECK, J., VANDECASTEELE, K., DE NEVE, W., DINGEMANS, A. M. C., EL NAQA, I., DEASY, J., BRADLEY, J., HUANG, E. & LAMBIN, P. 2010. Development, external validation and clinical usefulness of a practical prediction model for radiation-induced dysphagia in lung cancer patients. *Radiotherapy and Oncology,* 97**,** 455-461.

DELONG, E. R., DELONG, D. M. & CLARKE-PEARSON, D. L. 1988. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics,* 44**,** 837-845.

DERSIMONIAN, R. & LAIRD, N. 1986. Meta-analysis in clinical trials. *Controlled Clinical Trials,* 7**,** 177-188.

DHAMOON, M. S., TAI, W., BODEN-ALBALA, B., RUNDEK, T., PAIK, M. C., SACCO, R. L. & ELKIND, M. S. V. 2007. Risk of myocardial infarction or vascular death after first ischemic stroke: The northern Manhattan study. *Stroke,* 38**,** 1752-1758.

DIENER, H.-C., RINGLEB, P. A. & SAVI, P. 2005. Clopidogrel for the secondary prevention of stroke. *Expert Opinion on Pharmacotherapy,* 6**,** 755-764.

DOBSON, A. J. 1991. *An Introduction to generalized linear models*, Chapman & Hall.

DONNAN, G. A., FISHER, M., MACLEOD, M. & DAVIS, S. M. 2008. Stroke. *The Lancet,* 371**,** 1612-1623.

DORRESTEIJN, J. A. N., VISSEREN, F. L. J., RIDKER, P. M., PAYNTER, N. P., WASSINK, A. M. J., BURING, J. E., VAN DER GRAAF, Y. & COOK, N. R. 2011a. Aspirin for primary prevention of vascular events in women: individualized prediction of treatment effects. *European Heart Journal*.

DORRESTEIJN, J. A. N., VISSEREN, F. L. J., RIDKER, P. M., WASSINK, A. M. J., PAYNTER, N. P., STEYERBERG, E. W., VAN DER GRAAF, Y. & COOK, N. R. 2011b. Estimating treatment effects for individual patients based on the results of randomised clinical trials. *Bmj,* 343.

EASTON, J. D., SAVER, J. L., ALBERS, G. W., ALBERTS, M. J., CHATURVEDI, S., FELDMANN, E., HATSUKAMI, T. S., HIGASHIDA, R. T., JOHNSTON, S. C., KIDWELL, C. S., LUTSEP, H. L., MILLER, E. & SACCO, R. L. 2009. Definition and Evaluation of Transient Ischemic Attack: A Scientific Statement for Healthcare Professionals From the American Heart Association/American Stroke Association Stroke Council; Council on Cardiovascular Surgery and Anesthesia; Council on Cardiovascular Radiology and Intervention; Council on Cardiovascular Nursing; and the Interdisciplinary Council on Peripheral Vascular Disease: The American Academy of Neurology affirms the value of this statement as an educational tool for neurologists. *Stroke,* 40**,** 2276-2293.

EMBERSON, J., LEES, K. R., LYDEN, P., BLACKWELL, L., ALBERS, G., BLUHMKI, E., BROTT, T., COHEN, G., DAVIS, S., DONNAN, G., GROTTA, J., HOWARD, G., KASTE, M., KOGA, M., VON KUMMER, R., LANSBERG, M., LINDLEY, R. I., MURRAY, G., OLIVOT, J. M., PARSONS, M., TILLEY, B., TONI, D., TOYODA, K., WAHLGREN, N., WARDLAW, J., WHITELEY, W., DEL ZOPPO, G. J., BAIGENT, C., SANDERCOCK, P. & HACKE, W. 2014. Effect of treatment delay, age, and stroke severity on the effects of intravenous thrombolysis with alteplase for acute ischaemic stroke: a meta-analysis of individual patient data from randomised trials. *The Lancet*.

EVERS, S. M. A. A., STRUIJS, J. N., AMENT, A. J. H. A., VAN GENUGTEN, M. L. L., JAGER, J. C. & VAN DEN BOS, G. A. M. 2004. International Comparison of Stroke Cost Studies. *Stroke,* 35**,** 1209-1215.

FAROOQ, V., VAN KLAVEREN, D., STEYERBERG, E. W., MELIGA, E., VERGOUWE, Y., CHIEFFO, A., KAPPETEIN, A. P., COLOMBO, A., HOLMES JR, D. R., MACK, M., FELDMAN, T., MORICE, M.-C., STÅHLE, E., ONUMA, Y., MOREL, M.-A., GARCIA-GARCIA, H. M., VAN ES, G. A., DAWKINS, K. D., MOHR, F. W. & SERRUYS, P. W. 2013. Anatomical and clinical characteristics to guide decision making between coronary artery bypass surgery and percutaneous coronary intervention for individual patients: development and validation of SYNTAX score II. *The Lancet,* 381**,** 639-650.

FATOVICH, D. M. 2012. Believing is seeing: Stroke thrombolysis remains unproven after the third international stroke trial (IST-3). *Emergency Medicine Australasia,* 24**,** 477-479.

FATOVICH, D. M., MACDONALD, S. P. & BROWN, S. G. 2012. Thrombolysis in acute ischaemic stroke. *The Lancet,* 380**,** 1053.

FEWTRELL, M. S., KENNEDY, K., SINGHAL, A., MARTIN, R. M., NESS, A., HADDERS-ALGRA, M., KOLETZKO, B. & LUCAS, A. 2008. How much loss to follow-up is acceptable in long-term randomised trials and prospective studies? *Archives of Disease in Childhood,* 93**,** 458-461.

FINE, J. P. & GRAY, R. J. 1999. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association,* 94**,** 496-509.

FITZEK, S., LEISTRITZ, L., WITTE, O. W., HEUSCHMANN, P. U. & FITZEK, C. 2011. The Essen Stroke Risk Score in One-Year Follow-Up Acute Ischemic Stroke Patients. *Cerebrovascular Diseases,* 31**,** 400-407.

FLEISS, J. L. & COHEN, J. 1973. The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. *Educational and Psychological Measurement,* 33**,** 613-619.

FLINT, A. C., CULLEN, S. P., FAIGELES, B. S. & RAO, V. A. 2010. Predicting Long-Term Outcome after Endovascular Stroke Treatment: The Totaled Health Risks in Vascular Events Score. *American Journal of Neuroradiology,* 31**,** 1192-1196.

GELMAN, A. 2012. All Models are Right, Most are Useless. *Statistical Modeling, Causal Inference, and Social Science* [Online]. Available from: http://andrewgelman.com 2014].

GILLIGAN, A. K., THRIFT, A. G., STURM, J. W., DEWEY, H. M., MACDONELL, R. A. L. & DONNAN, G. A. 2005. Stroke Units, Tissue Plasminogen Activator, Aspirin and Neuroprotection: Which Stroke Intervention Could Provide the Greatest Community Benefit? *Cerebrovascular Diseases,* 20**,** 239-244.

GLASS, G. V. 1976. Primary, Secondary, and Meta-Analysis of Research. *Educational Researcher,* 5**,** 3-8.

GLASZIOU, P. & IRWIG, L. 1995. An evidence based approach to individualising treatment. *BMJ.*

GO, A. S., MOZAFFARIAN, D., ROGER, V. L., BENJAMIN, E. J., BERRY, J. D., BLAHA, M. J., DAI, S., FORD, E. S., FOX, C. S., FRANCO, S., FULLERTON, H. J., GILLESPIE, C., HAILPERN, S. M., HEIT, J. A., HOWARD, V. J., HUFFMAN, M. D., JUDD, S. E., KISSELA, B. M.,

KITTNER, S. J., LACKLAND, D. T., LICHTMAN, J. H., LISABETH, L. D., MACKEY, R. H., MAGID, D. J., MARCUS, G. M., MARELLI, A., MATCHAR, D. B., MCGUIRE, D. K., MOHLER, E. R., MOY, C. S., MUSSOLINO, M. E., NEUMAR, R. W., NICHOL, G., PANDEY, D. K., PAYNTER, N. P., REEVES, M. J., SORLIE, P. D., STEIN, J., TOWFIGHI, A., TURAN, T. N., VIRANI, S. S., WONG, N. D., WOO, D. & TURNER, M. B. 2014. Heart Disease and Stroke Statistics—2014 Update: A Report From the American Heart Association. *Circulation,* 129**,** e28-e292.

GOLDIE, F. C., FULTON, R. L., FRANK, B., LEES, K. R. & COLLABORATION, V. 2014. Interdependence of stroke outcome scales: reliable estimates from the Virtual International Stroke Trials Archive (VISTA). *International Journal of Stroke,* 9**,** 328-332.

GRAHAM, P. & JACKSON, R. 1993. The analysis of ordinal agreement data: beyond weighted kappa. *Journal of Clinical Epidemiology,* 46**,** 1055-1062.

GRAMBSCH, P. M. & THERNEAU, T. M. 1994. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika,* 81**,** 515-526.

GRAY, B. 2013. cmprsk: Subdistribution Analysis of Competing Risks. *R package version 2.2-6*.

GRAY, L. J., ALI, M., LYDEN, P. D. & BATH, P. M. W. 2009. Interconversion of the National Institutes of Health Stroke Scale and Scandinavian Stroke Scale in Acute Stroke. *Journal of stroke and cerebrovascular diseases : the official journal of National Stroke Association,* 18**,** 466-468.

GREENWOOD, M. 1926. *A Report on the Natural Duration of Cancer*, H.M. Stationery Office.

GROVE, W. M., ZALD, D. H., LEBOW, B. S., SNITZ, B. E. & NELSON, C. 2000. Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment,* 12**,** 19.

GROVES, T. & GODLEE, F. 2012. Open science and reproducible research. *Bmj,* 344.

GRUBE, M. M., KOENNECKE, H.-C., WALTER, G., MEISEL, A., SOBESKY, J., NOLTE, C. H., WELLWOOD, I., HEUSCHMANN, P. U. & ON BEHALF OF THE BERLIN STROKE, R. 2013. Influence of Acute Complications on Outcome 3 Months after Ischemic Stroke. *PLoS ONE,* 8**,** e75719.

GUYATT, G., CAIRNS, J., CHURCHILL, D. & ET AL. 1992. Evidence-based medicine: A new approach to teaching the practice of medicine. *JAMA,* 268**,** 2420-2425.

HACKE, W., KASTE, M., BLUHMKI, E., BROZMAN, M., DÁVALOS, A., GUIDETTI, D., LARRUE, V., LEES, K. R., MEDEGHRI, Z., MACHNIG,

T., SCHNEIDER, D., VON KUMMER, R., WAHLGREN, N. & TONI, D. 2008. Thrombolysis with Alteplase 3 to 4.5 Hours after Acute Ischemic Stroke. *New England Journal of Medicine,* 359**,** 1317-1329.

HACKE, W., KASTE, M., FIESCHI, C., VON KUMMER, R., DAVALOS, A., MEIER, D., LARRUE, V., BLUHMKI, E., DAVIS, S., DONNAN, G., SCHNEIDER, D., DIEZ-TEJEDOR, E. & TROUILLAS, P. 1998. Randomised double-blind placebo-controlled trial of thrombolytic therapy with intravenous alteplase in acute ischaemic stroke (ECASS II). *The Lancet,* 352**,** 1245-1251.

HACKETT, M. L., YAPA, C., PARAG, V. & ANDERSON, C. S. 2005. Frequency of Depression After Stroke: A Systematic Review of Observational Studies. *Stroke,* 36**,** 1330-1340.

HARRELL, F. E. 2001. *Regression Modeling Strategies: With applications to linear models, logistic regression, and survival analysis*, Springer.

HARRELL, F. E. 2013. rms: Regression Modeling Strategies. R package version 3.6-3. *http://CRAN.R-project.org/package=rms*.

HARRELL, F. E., LEE, K. L. & MARK, D. B. 1996. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine,* 15**,** 361-387.

HARRELL, F. E., LEE, K. L. & POLLOCK, B. G. 1988. Regression Models in Clinical Studies: Determining Relationships Between Predictors and Response. *Journal of the National Cancer Institute,* 80**,** 1198-1202.

HARRELL, F. E., MARGOLIS, P. A., GOVE, S., MASON, K. E., MULHOLLAND, E. K., LEHMANN, D., MUHE, L., GATCHALIAN, S. & EICHENWALD, H. F. 1998. Development of a clinical prediction model for an ordinal outcome: the World Health Organization Multicentre Study of Clinical Signs and Etiological Agents of Pneumonia, Sepsis and Meningitis in Young Infants. *Statistics in Medicine,* 17**,** 909-944.

HATANO, S. 1976. Experience from a multicentre stroke register: a preliminary report. *Bulletin of the World Health Organization,* 54**,** 541.

HAYDEN, J. A., CÔTÉ, P. & BOMBARDIER, C. 2006. Evaluation of the Quality of Prognosis Studies in Systematic Reviews. *Annals of Internal Medicine,* 144**,** 427-437.

HAYWARD, R., KENT, D., VIJAN, S. & HOFER, T. 2006. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. *BMC Medical Research Methodology,* 6**,** 18.

HEMINGWAY, H., CROFT, P., PEREL, P., HAYDEN, J. A., ABRAMS, K., TIMMIS, A., BRIGGS, A., UDUMYAN, R., MOONS, K. G. M.,

STEYERBERG, E. W., ROBERTS, I., SCHROTER, S., ALTMAN, D. G. & RILEY, R. D. 2013. *Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes*.

HIGGINS, J. P. T., THOMPSON, S. G. & SPIEGELHALTER, D. J. 2009. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 172**,** 137-159.

HILL, M. D. & COUTTS, S. B. 2014. Alteplase in acute ischaemic stroke: the need for speed. *The Lancet*.

HILLEN, T., COSHALL, C., TILLING, K., RUDD, A. G., MCGOVERN, R. & WOLFE, C. D. A. 2003. Cause of Stroke Recurrence Is Multifactorial: Patterns, Risk Factors, and Outcomes of Stroke Recurrence in the South London Stroke Register. *Stroke,* 34**,** 1457-1463.

HINGORANI, A. D., WINDT, D. A. V. D., RILEY, R. D., ABRAMS, K., MOONS, K. G. M., STEYERBERG, E. W., SCHROTER, S., SAUERBREI, W., ALTMAN, D. G. & HEMINGWAY, H. 2013. *Prognosis research strategy (PROGRESS) 4: Stratified medicine research*.

HOFFMAN, J. R. & COOPER, R. J. 2012. How is more negative evidence being used to support claims of benefit: The curious case of the third international stroke trial (IST-3). *Emergency Medicine Australasia,* 24**,** 473-476.

HONG, K. S., KANG, D. W., KOO, J. S., YU, K. H., HAN, M. K., CHO, Y. J., PARK, J. M., BAE, H. J. & LEE, B. C. 2008. Impact of neurological and medical complications on 3-month outcomes in acute ischaemic stroke. *European Journal of Neurology,* 15**,** 1324-1331.

HOSMER JR, D. W., LEMESHOW, S. & MAY, S. 2011. *Applied survival analysis: regression modeling of time to event data*, Wiley. com.

HUDIS, C. A. 2007. Trastuzumab — Mechanism of Action and Use in Clinical Practice. *New England Journal of Medicine,* 357**,** 39-51.

IST COLLABORATIVE GROUP 1997. The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19435 patients with acute ischaemic stroke. *The Lancet,* 349**,** 1569-1581.

JACKSON, C., CROSSLAND, L., DENNIS, M., WARDLAW, J. & SUDLOW, C. 2008. Assessing the impact of the requirement for explicit consent in a hospital-based stroke study. *QJM,* 101**,** 281-289.

JACKSON, C. A., HUTCHISON, A., DENNIS, M. S., WARDLAW, J. M., LEWIS, S. C. & SUDLOW, C. L. M. 2009. Differences Between Ischemic Stroke Subtypes in Vascular Outcomes Support a Distinct Lacunar Ischemic Stroke Arteriopathy: A Prospective, Hospital-Based Study. *Stroke,* 40**,** 3679-3684.

JACKSON, C. A., HUTCHISON, A., DENNIS, M. S., WARDLAW, J. M., LINDGREN, A., NORRVING, B., ANDERSON, C. S., HANKEY, G. J., JAMROZIK, K., APPELROS, P. & SUDLOW, C. L. M. 2010. Differing Risk Factor Profiles of Ischemic Stroke Subtypes: Evidence for a Distinct Lacunar Arteriopathy? *Stroke,* 41**,** 624-629.

JANSSEN, K., VERGOUWE, Y., KALKMAN, C., GROBBEE, D. & MOONS, K. 2009. A simple method to adjust clinical prediction models to local circumstances. *Canadian Journal of Anesthesia / Journal canadien d'anesthésie,* 56**,** 194-201.

JOHNSTON, S. C., ROTHWELL, P. M., NGUYEN-HUYNH, M. N., GILES, M. F., ELKINS, J. S., BERNSTEIN, A. L. & SIDNEY, S. 2007. Validation and refinement of scores to predict very early stroke risk after transient ischaemic attack. *The Lancet,* 369**,** 283-292.

KAHNEMAN, D. 2011. *Thinking, fast and slow*, Macmillan.

KAHNEMAN, D. & KLEIN, G. 2009. Conditions for intuitive expertise: A failure to disagree. *American Psychologist,* 64**,** 515-526.

KAMOUCHI, M., KUMAGAI, N., OKADA, Y., ORIGASA, H., YAMAGUCHI, T. & KITAZONO, T. 2012. Risk Score for Predicting Recurrence in Patients with Ischemic Stroke: The Fukuoka Stroke Risk Score for Japanese. *Cerebrovascular Diseases,* 34**,** 351-357.

KAPLAN, E. L. & MEIER, P. 1958. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association,* 53**,** 457-481.

KATTAN, M. W. & GERDS, T. A. 2011. Making and evaluating a statistical prediction model for the absolute risk of prostate cancer recurrence. *Cancer,* 117**,** 5026-5028.

KENNETH, H. & GENTLEMAN, R. 2010. muhaz: Hazard Function Estimation in Survival Analysis.

KENT, D. & HAYWARD, R. 2007a. When Averages Hide Individual Differences in Clinical Trials. *American Scientist*.

KENT, D., ROTHWELL, P., IOANNIDIS, J., ALTMAN, D. & HAYWARD, R. 2010. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials,* 11**,** 85.

KENT, D. M. & HAYWARD, R. A. 2007b. Limitations of applying summary results of clinical trials to individual patients: The need for risk stratification. *JAMA,* 298**,** 1209-1212.

KENT, D. M., SELKER, H. P., RUTHAZER, R., BLUHMKI, E. & HACKE, W. 2006. The Stroke–Thrombolytic Predictive Instrument: A Predictive

Instrument for Intravenous Thrombolysis in Acute Ischemic Stroke. *Stroke,* 37**,** 2957-2962.

KERN, S. E. 2012. Why Your New Cancer Biomarker May Never Work: Recurrent Patterns and Remarkable Diversity in Biomarker Failures. *Cancer Research,* 72**,** 6097-6101.

KERNAN, W. N., HORWITZ, R. I., BRASS, L. M., VISCOLI, C. M. & TAYLOR, K. J. W. 1991. A Prognostic System for Transient Ischemia or Minor Stroke. *Annals of Internal Medicine,* 114**,** 552-557.

KERNAN, W. N., VISCOLI, C. M., BRASS, L. M., MAKUCH, R. W., SARREL, P. M., ROBERTS, R. S., GENT, M., ROTHWELL, P., SACCO, R. L., LIU, R. C., BODEN-ALBALA, B. & HORWITZ, R. I. 2000. The stroke prognosis instrument II (SPI-II): A clinical prediction instrument for patients with transient ischemia and nondisabling ischemic stroke. *Stroke,* 31**,** 456-462.

KERR, K. F., WANG, Z., JANES, H., MCCLELLAND, R. L., PSATY, B. M. & PEPE, M. S. 2014. Net reclassification indices for evaluating risk prediction instruments: a critical review. *Epidemiology,* 25**,** 114-121.

KOLLER, M. T., RAATZ, H., STEYERBERG, E. W. & WOLBERS, M. 2012. Competing risks and the clinical community: irrelevance or ignorance? *Statistics in Medicine,* 31**,** 1089-1097.

KÖNIG, I. R., MALLEY, J. D., WEIMAR, C., DIENER, H. C. & ZIEGLER, A. 2007. Practical experiences on the necessity of external validation. *Statistics in Medicine,* 26**,** 5499-5511.

KÖNIG, I. R., ZIEGLER, A., BLUHMKI, E., HACKE, W., BATH, P. M. W., SACCO, R. L., DIENER, H. C., WEIMAR, C. & INVESTIGATORS, O. B. O. T. V. I. S. T. A. 2008. Predicting Long-Term Outcome After Acute Ischemic Stroke: A Simple Index Works in Patients From Controlled Clinical Trials. *Stroke,* 39**,** 1821-1826.

KWAKKEL, G., WAGENAAR, R. C., KOLLEN, B. J. & LANKHORST, G. J. 1996. Predicting Disability in Stroke—A Critical Review of the Literature. *Age and Ageing 1996:25:479-489*.

LANGHORNE, P., STOTT, D. J., ROBERTSON, L., MACDONALD, J., JONES, L., MCALPINE, C., DICK, F., TAYLOR, G. S. & MURRAY, G. 2000. Medical Complications After Stroke: A Multicenter Study. *Stroke,* 31**,** 1223-1229.

LAU, B., COLE, S. R. & GANGE, S. J. 2009. Competing Risk Regression Models for Epidemiologic Data. *American Journal of Epidemiology,* 170**,** 244-256.

LAUPACIS, A., SEKAR, N. & STIELL, L. G. 1997. Clinical Prediction Rules. *JAMA: The Journal of the American Medical Association,* 277**,** 488-494.

LEE, Y.-S., CHEN, D.-Y., CHEN, Y.-M., CHUANG, Y.-W., LIAO, S.-C., LIN, C.-S., TANG, Y.-J., TSAI, J.-J., LAN, J.-L. & HSU, H.-Y. 2009. First-ever ischemic stroke in Taiwanese elderly patients: predicting functional independence after a 6-month follow-up. *Archives of Gerontology and Geriatrics,* 49, Supplement 2**,** S26-S31.

LEES, K. R., BLUHMKI, E., VON KUMMER, R., BROTT, T. G., TONI, D., GROTTA, J. C., ALBERS, G. W., KASTE, M., MARLER, J. R., HAMILTON, S. A., TILLEY, B. C., DAVIS, S. M., DONNAN, G. A. & HACKE, W. 2010. Time to treatment with intravenous alteplase and outcome in stroke: an updated pooled analysis of ECASS, ATLANTIS, NINDS, and EPITHET trials. *The Lancet,* 375**,** 1695-1703.

LEMESHOW, S. & HOSMER, D. W. 1982. A review of goodness of fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology,* 115**,** 92-106.

LEMMENS, R., SMET, S. & THIJS, V. N. 2013. Clinical Scores for Predicting Recurrence After Transient Ischemic Attack or Stroke: How Good are They? *Stroke,* 44**,** 1198-1203.

LI, W.-J., GAO, Z.-Y., HE, Y., LIU, G.-Z. & GAO, X.-G. 2012. Application and Performance of Two Stroke Outcome Prediction Models in a Chinese Population. *PM&R,* 4**,** 123-128.

LITTLE, R. J. A. 1992. Regression With Missing X's: A Review. *Journal of the American Statistical Association,* 87**,** 1227-1237.

LOU, M., SAFDAR, A., MEHDIRATTA, M., KUMAR, S., SCHLAUG, G., CAPLAN, L., SEARLS, D. & SELIM, M. 2008. The HAT Score: A simple grading scale for predicting hemorrhage after thrombolysis. *Neurology,* 71**,** 1417-1423.

LYDEN, P., LU, M., JACKSON, C., MARLER, J., KOTHARI, R., BROTT, T. & ZIVIN, J. 1999. Underlying Structure of the National Institutes of Health Stroke Scale: Results of a Factor Analysis. *Stroke,* 30**,** 2347-2354.

LYDEN, P. D. 2012. In Anticipation of International Stroke Trial-3 (IST-3). *Stroke,* 43**,** 1691-1694.

MAAS, A. I. R., MURRAY, G. D., ROOZENBEEK, B., LINGSMA, H. F., BUTCHER, I., MCHUGH, G. S., WEIR, J., LU, J. & STEYERBERG, E. W. 2013. Advancing care for traumatic brain injury: findings from the IMPACT studies and perspectives on future research. *The Lancet Neurology,* 12**,** 1200-1210.

MAHONEY, F. & BARTHEL, D. 1965. Functional evaluation: the Barthel Index. *Maryland State Medical Journal,* 14**,** 61-65.

MAIER, I. L., BAUERLE, M., KERMER, P., HELMS, H. J. & BUETTNER, T. 2013. Risk prediction of very early recurrence, death and progression after acute ischaemic stroke. *European Journal of Neurology,* 20**,** 599-604.

MALLETT, S., ROYSTON, P., DUTTON, S., WATERS, R. & ALTMAN, D. 2010. Reporting methods in studies developing prognostic models in cancer: a review. *BMC Medicine,* 8**,** 20.

MALOTTKI, K., BISWAS, M., DEEKS, J. J., RILEY, R. D., CRADDOCK, C., JOHNSON, P. & BILLINGHAM, L. 2014. Stratified medicine in European Medicines Agency licensing: a systematic review of predictive biomarkers. *BMJ Open,* 4.

MARKS, R. J., SIMONS, R. S., BLIZZARD, R. A. & BROWNE, D. R. G. 1991. Predicting outcome in intensive therapy units —a comparison of Apache II with subjective assessments. *Intensive Care Medicine,* 17**,** 159-163.

MAZYA, M., EGIDO, J. A., FORD, G. A., LEES, K. R., MIKULIK, R., TONI, D., WAHLGREN, N. & AHMED, N. 2012. Predicting the Risk of Symptomatic Intracerebral Hemorrhage in Ischemic Stroke Treated With Intravenous Alteplase: Safe Implementation of Treatments in Stroke (SITS) Symptomatic Intracerebral Hemorrhage Risk Score. *Stroke,* 43**,** 1524-1531.

MAZYA, M. V., BOVI, P., CASTILLO, J., JATUZIS, D., KOBAYASHI, A., WAHLGREN, N. & AHMED, N. 2013. External Validation of the SEDAN Score for Prediction of Intracerebral Hemorrhage in Stroke Thrombolysis. *Stroke,* 44**,** 1595-1600.

MCCULLAGH, P. 1980. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society. Series B (Methodological),* 42**,** 109-142.

MCHUGH, G. S., BUTCHER, I., STEYERBERG, E. W., MARMAROU, A., LU, J., LINGSMA, H. F., WEIR, J., MAAS, A. I. R. & MURRAY, G. D. 2010. A simulation study evaluating approaches to the analysis of ordinal outcome data in randomized controlled trials in traumatic brain injury: results from the IMPACT Project. *Clinical Trials,* 7**,** 44-57.

MCMEEKIN, P., FLYNN, D., FORD, G. A., RODGERS, H. & THOMSON, R. G. 2012. Validating the Stroke-Thrombolytic Predictive Instrument in a Population in the United Kingdom. *Stroke,* 43**,** 3378-3381.

MEADS, C., AHMED, I. & RILEY, R. 2012. A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. *Breast Cancer Research and Treatment,* 132**,** 365-377.

MEEHL, P. E. 1954. *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence,* Minneapolis, MN, US, University of Minnesota Press.

MENG, X., WANG, Y., ZHAO, X., WANG, C., LI, H., LIU, L., ZHOU, Y., XU, J. & WANG, Y. 2011. Validation of the Essen Stroke Risk Score and the Stroke Prognosis Instrument II in Chinese Patients. *Stroke,* 42**,** 3619-3620.

MENON, B. K., SAVER, J. L., PRABHAKARAN, S., REEVES, M., LIANG, L., OLSON, D. M., PETERSON, E. D., HERNANDEZ, A. F., FONAROW, G. C., SCHWAMM, L. H. & SMITH, E. E. 2012. Risk Score for Intracranial Hemorrhage in Patients With Acute Ischemic Stroke Treated With Intravenous Tissue-Type Plasminogen Activator. *Stroke,* 43**,** 2293-2299.

MOHAN, K. M., CRICHTON, S. L., GRIEVE, A. P., RUDD, A. G., WOLFE, C. D. A. & HEUSCHMANN, P. U. 2009. Frequency and predictors for the risk of stroke recurrence up to 10 years after stroke: the South London Stroke Register. *Journal of Neurology, Neurosurgery & Psychiatry,* 80**,** 1012-1018.

MOHAN, K. M., WOLFE, C. D. A., RUDD, A. G., HEUSCHMANN, P. U., KOLOMINSKY-RABAS, P. L. & GRIEVE, A. P. 2011. Risk and Cumulative Risk of Stroke Recurrence: a Systematic Review and Meta-Analysis. *Stroke,* 42**,** 1489-1494.

MOHR, J. P., ALBERS, G. W., AMARENCO, P., BABIKIAN, V. L., BILLER, J., BREY, R. L., COULL, B., EASTON, J. D., GOMEZ, C. R., HELGASON, C. M., KASE, C. S., PULLICINO, P. M. & TURPIE, A. G. G. 1997. Etiology of Stroke. *Stroke,* 28**,** 1501-1506.

MOONS, K. G. M., DONDERS, R. A. R. T., STIJNEN, T. & HARRELL JR, F. E. 2006. Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology,* 59**,** 1092-1101.

MOONS, K. G. M., KENGNE, A. P., GROBBEE, D. E., ROYSTON, P., VERGOUWE, Y., ALTMAN, D. G. & WOODWARD, M. 2012a. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart,* 98**,** 691-698.

MOONS, K. G. M., KENGNE, A. P., WOODWARD, M., ROYSTON, P., VERGOUWE, Y., ALTMAN, D. G. & GROBBEE, D. E. 2012b. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart,* 98**,** 683-690.

MOONS, K. G. M., ROYSTON, P., VERGOUWE, Y., GROBBEE, D. E. & ALTMAN, D. G. 2009. Prognosis and prognostic research: what, why, and how? *Bmj,* 338.

MULTICENTRE ACUTE STROKE TRIAL - ITALY, G. 1995. Randomised controlled trial of streptokinase, aspirin, and combination of both in treatment of acute ischaemic stroke. *The Lancet,* 346**,** 1509-1514.

MUSCARI, A., PUDDU, G. M., SANTORO, N. & ZOLI, M. 2011. A simple scoring system for outcome prediction of ischemic stroke. *Acta Neurologica Scandinavica,* 124**,** 334-342.

MUSHKUDIANI, N. A., HUKKELHOVEN, C. W. P. M., HERNÁNDEZ, A. V., MURRAY, G. D., CHOI, S. C., MAAS, A. I. R. & STEYERBERG, E. W. 2008. A systematic review finds methodological improvements necessary for prognostic models in determining traumatic brain injury outcomes. *Journal of Clinical Epidemiology,* 61**,** 331-343.

NAGELKERKE, N. J. D. 1991. A note on a general definition of the coefficient of determination. *Biometrika,* 78**,** 691-692.

NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE (NICE). 2006. *Statins for the prevention of cardiovascular events* [Online]. Available: guidance.nice.org.uk/ta94 [Accessed 27/06 2014].

NAVI, B. B., KAMEL, H., SIDNEY, S., KLINGMAN, J. G., NGUYEN-HUYNH, M. N. & JOHNSTON, S. C. 2011. Validation of the stroke Prognostic Instrument-II in a large, modern, community-based cohort of ischemic stroke survivors. *Stroke,* 42**,** 3392-3396.

NEW, P. W. & BUCHBINDER, R. 2006. Critical Appraisal and Review of the Rankin Scale and Its Derivatives. *Neuroepidemiology,* 26**,** 4-15.

NEWCOMBE, R. G. 1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine,* 17**,** 857-872.

NTAIOS, G., FAOUZI, M., FERRARI, J., LANG, W., VEMMOS, K. & MICHEL, P. 2012. An integer-based score to predict functional outcome in acute ischemic stroke: The ASTRAL score. *Neurology,* 78**,** 1916-1922.

NÚÑEZ, E., STEYERBERG, E. W. & NÚÑEZ, J. 2011. Regression Modeling Strategies. *Revista Española de Cardiología,* 64**,** 501-507.

PAIKIN, J. S. & EIKELBOOM, J. W. 2012. Aspirin. *Circulation,* 125**,** e439-e442.

PAPAVASILEIOU, V., MILIONIS, H., MICHEL, P., MAKARITSIS, K., VEMMOU, A., KOROBOKI, E., MANIOS, E., VEMMOS, K. & NTAIOS, G. 2013. ASTRAL Score Predicts 5-Year Dependence and Mortality in Acute Ischemic Stroke. *Stroke,* 44**,** 1616-1620.

PEDUZZI, P., CONCATO, J., KEMPER, E., HOLFORD, T. R. & FEINSTEIN, A. R. 1996. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology,* 49**,** 1373-1379.

PENCINA, M. J., D' AGOSTINO, R. B. & VASAN, R. S. 2008. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medicine,* 27**,** 157-172.

PENCINA, M. J., D'AGOSTINO, R. B. & STEYERBERG, E. W. 2011. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in Medicine,* 30**,** 11-21.

PEREL, P., EDWARDS, P., WENTZ, R. & ROBERTS, I. 2006. Systematic review of prognostic models in traumatic brain injury. *BMC Medical Informatics & Decision Making,* 6**,** 38-10.

PETRELLA, J. R. & PROVENZALE, J. M. 2000. MR Perfusion Imaging of the Brain. *American Journal of Roentgenology,* 175**,** 207-219.

PEXMAN, J. H. W., BARBER, P. A., HILL, M. D., SEVICK, R. J., DEMCHUK, A. M., HUDON, M. E., HU, W. Y. & BUCHAN, A. M. 2001. Use of the Alberta Stroke Program Early CT Score (ASPECTS) for Assessing CT Scans in Patients with Acute Stroke. *American Journal of Neuroradiology,* 22**,** 1534-1542.

PEZZINI, A., GRASSI, M., DEL ZOTTO, E., LODIGIANI, C., FERRAZZI, P., SPALLONI, A., PATELLA, R., GIOSSI, A., VOLONGHI, I., IACOVIELLO, L., MAGONI, M., ROTA, L. L., RASURA, M. & PADOVANI, A. 2009. Common genetic markers and prediction of recurrent events after ischemic stroke in young adults. *Neurology,* 73**,** 717-723.

PINCE, J. 1981. MD Thesis, Université Paul Sabatier.

POCOCK, S. J., ASSMANN, S. E., ENOS, L. E. & KASTEN, L. E. 2002. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. *Statistics in Medicine,* 21**,** 2917-2930.

POCOCK, S. J., MEHRAN, R., CLAYTON, T. C., NIKOLSKY, E., PARISE, H., FAHY, M., LANSKY, A. J., BERTRAND, M. E., LINCOFF, A. M., MOSES, J. W., OHMAN, E. M., WHITE, H. D. & STONE, G. W. 2010. Prognostic Modeling of Individual Patient Risk and Mortality Impact of Ischemic and Hemorrhagic Complications: Assessment From the Acute Catheterization and Urgent Intervention Triage Strategy Trial. *Circulation,* 121**,** 43-51.

POCOCK, S. J., STONE, G. W., MEHRAN, R. & CLAYTON, T. C. 2014. Individualizing treatment choices using quantitative methods. *American Heart Journal*.

PONGMORAGOT, J., RABINSTEIN, A. A., NILANONT, Y., SWARTZ, R. H., ZHOU, L., SAPOSNIK, G. & THE INVESTIGATORS OF THE REGISTRY OF THE CANADIAN STROKE NETWORK UNIVERSITY OF TORONTO STROKE PROGRAM FOR THE STROKE OUTCOMES RESEARCH CANADA WORKING GROUP 2013. Pulmonary Embolism in Ischemic

Stroke: Clinical Presentation, Risk Factors, and Outcome. *Journal of the American Heart Association,* 2.

PUETZ, V., DZIALOWSKI, I., HILL, M. D. & DEMCHUK, A. M. 2009. The Alberta Stroke Program Early CT Score in clinical practice: what have we learned? *International Journal of Stroke,* 4**,** 354-364.

PUTAALA, J., HAAPANIEMI, E., METSO, A. J., METSO, T. M., ARTTO, V., KASTE, M. & TATLISUMAK, T. 2010. Recurrent ischemic events in young adults after first-ever ischemic stroke. *Annals of Neurology,* 68**,** 661-671.

PUTTER, H., FIOCCO, M. & GESKUS, R. B. 2007. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine,* 26**,** 2389-2430.

PUTTER, H., SASAKO, M., HARTGRINK, H. H., VAN DE VELDE, C. J. H. & VAN HOUWELINGEN, J. C. 2005. Long-term survival with non-proportional hazards: results from the Dutch Gastric Cancer Trial. *Statistics in Medicine,* 24**,** 2807-2821.

QUINN, T. J., DAWSON, J., WALTERS, M. R. & LEES, K. R. 2009. Functional outcome measures in contemporary stroke trials. *International Journal of Stroke,* 4**,** 200-205.

RATHORE, S. S., HINN, A. R., COOPER, L. S., TYROLER, H. A. & ROSAMOND, W. D. 2002. Characterization of Incident Stroke Signs and Symptoms: Findings From the Atherosclerosis Risk in Communities Study. *Stroke,* 33**,** 2718-2721.

REID, J. M., GUBITZ, G. J., DAI, D., KYDD, D., ESKES, G., REIDY, Y., CHRISTIAN, C., COUNSELL, C. E., DENNIS, M. & PHILLIPS, S. J. 2010. Predicting functional outcome after stroke by modelling baseline clinical and CT variables. *Age and Ageing,* 39**,** 360-366.

RESCHE-RIGON, M., WHITE, I. R., BARTLETT, J. W., PETERS, S. A. E., THOMPSON, S. G. & ON BEHALF OF THE, P.-I. M. T. S. G. 2013. Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. *Statistics in Medicine,* 32**,** 4890-4905.

RILEY, R. D., HAYDEN, J. A., STEYERBERG, E. W., MOONS, K. G. M., ABRAMS, K., KYZAS, P. A., MALATS, N., BRIGGS, A., SCHROTER, S., ALTMAN, D. G., HEMINGWAY, H. & FOR THE PROGRESS GROUP 2013. Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. *PLoS Med,* 10**,** e1001380.

RILEY, R. D., HIGGINS, J. P. T. & DEEKS, J. J. 2011. Interpretation of random effects meta-analyses. *Bmj,* 342.

RILEY, R. D., LAMBERT, P. C. & ABO-ZAID, G. 2010. *Meta-analysis of individual participant data: rationale, conduct, and reporting*.

ROBIN, X., TURCK, N., HAINARD, A., TIBERTI, N., LISACEK, F., SANCHEZ, J.-C. & MULLER, M. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics,* 12**,** 77.

ROBINSON, L. D. & JEWELL, N. P. 1991. Some Surprising Results about Covariate Adjustment in Logistic Regression Models. *International Statistical Review / Revue Internationale de Statistique,* 59**,** 227-240.

RÖDÉN-JÜLLIG, Å., BRITTON, M., MALMKVIST, K. & LEIJD, B. 2003. Aspirin in the prevention of progressing stroke: a randomized controlled study. *Journal of Internal Medicine,* 254**,** 584-590.

ROTHWELL, P. M. 2007a. *Treating individuals: from randomised trials to personalised medicine*, Elsevier Health Sciences.

ROTHWELL, P. M. 2007b. When should we expect clinically important differences in response to treatment? *Treating Individuals: From Randomised Trials to Personalised Medicine***,** 139-150.

ROYSTON, P. & ALTMAN, D. 2013. External validation of a Cox prognostic model: principles and methods. *BMC Medical Research Methodology,* 13**,** 33.

ROYSTON, P. & ALTMAN, D. G. 1994. Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* 43**,** 429-467.

ROYSTON, P., AMBLER, G. & SAUERBREI, W. 1999. The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology,* 28**,** 964-974.

ROYSTON, P., MOONS, K. G. M., ALTMAN, D. G. & VERGOUWE, Y. 2009. Prognosis and prognostic research: Developing a prognostic model. *Bmj,* 338.

RUBIN, D. B. 1996. Multiple Imputation After 18+ Years. *Journal of the American Statistical Association,* 91**,** 473-489.

RUIGROK, Y. M., BUSKENS, E. & RINKEL, G. J. E. 2001. Attributable Risk of Common and Rare Determinants of Subarachnoid Hemorrhage. *Stroke,* 32**,** 1173-1175.

SACCO, R. L., BENJAMIN, E. J., BRODERICK, J. P., DYKEN, M., EASTON, J. D., FEINBERG, W. M., GOLDSTEIN, L. B., GORELICK, P. B., HOWARD, G., KITTNER, S. J., MANOLIO, T. A., WHISNANT, J. P. & WOLF, P. A. 1997. Risk Factors. *Stroke,* 28**,** 1507-1517.

SACCO, R. L., KASNER, S. E., BRODERICK, J. P., CAPLAN, L. R., CONNORS, J. J., CULEBRAS, A., ELKIND, M. S. V., GEORGE, M. G., HAMDAN, A. D., HIGASHIDA, R. T., HOH, B. L., JANIS, L. S., KASE, C. S., KLEINDORFER, D. O., LEE, J.-M., MOSELEY, M. E., PETERSON, E. D., TURAN, T. N., VALDERRAMA, A. L. & VINTERS, H. V. 2013. An Updated Definition of Stroke for the 21st Century: A Statement for Healthcare Professionals From the American Heart Association/American Stroke Association. *Stroke,* 44**,** 2064-2089.

SAKA, Ö., MCGUIRE, A. & WOLFE, C. 2009. Cost of stroke in the United Kingdom. *Age and Ageing,* 38**,** 27-32.

SALISBURY, A. C., WANG, K., COHEN, D. J., LI, Y., JONES, P. G. & SPERTUS, J. A. 2013. Selecting Antiplatelet Therapy at the Time of Percutaneous Intervention for an Acute Coronary Syndrome: Weighing the Benefits and Risks of Prasugrel Versus Clopidogrel. *Circulation: Cardiovascular Quality and Outcomes,* 6**,** 27-34.

SALZINGER, K. 2005. Clinical, Statistical, and Broken-Leg Predictions. *Behavior and Philosophy,* 33**,** 91-99.

SANDERCOCK, P., COUNSELL, C., TSENG, M. & CECCONI, E. 2014. Oral antiplatelet therapy for acute ischaemic stroke. *Cochrane Database Syst Rev.,* 3:CD000029.**,** 10.1002/14651858.CD000029.pub3.

SANDERCOCK, P., LINDLEY, R., WARDLAW, J., DENNIS, M., INNES, K., COHEN, G., WHITELEY, W., PERRY, D., SOOSAY, V., BUCHANAN, D., VENABLES, G., CZLONKOWSKA, A., KOBAYASHI, A., BERGE, E., SLOT, K., MURRAY, V., PEETERS, A., HANKEY, G., MATZ, K., BRAININ, M., RICCI, S., CANTISANI, T., GUBITZ, G., PHILLIPS, S., ANTONIO, A., CORREIA, M., LYRER, P., KANE, I., LUNDSTROM, E. & GROUP, T. I.-C. 2011. Update on the third international stroke trial (IST-3) of thrombolysis for acute ischaemic stroke and baseline features of the 3035 patients recruited. *Trials,* 12**,** 252.

SANDERCOCK, P. A. G., COUNSELL, C., GUBITZ, G. J. & TSENG, M.-C. 2008. Antiplatelet therapy for acute ischaemic stroke. *Cochrane Database of Systematic Reviews* [Online].

SAPOSNIK, G., COTE, R., MAMDANI, M., RAPTIS, S., THORPE, K. E., FANG, J., REDELMEIER, D. A. & GOLDSTEIN, L. B. 2013a. JURaSSiC: Accuracy of clinician vs risk score prediction of ischemic stroke outcomes. *Neurology,* 81**,** 448-455.

SAPOSNIK, G., GUZIK, A. K., REEVES, M., OVBIAGELE, B. & JOHNSTON, S. C. 2013b. Stroke Prognostication using Age and NIH Stroke Scale: SPAN-100. *Neurology,* 80**,** 21-28.

SAPOSNIK, G., KAPRAL, M. K., LIU, Y., HALL, R., O'DONNELL, M., RAPTIS, S., TU, J. V., MAMDANI, M. & AUSTIN, P. C. 2011. IScore: A Risk Score to Predict Death Early After Hospitalization for an Acute Ischemic Stroke. *Circulation,* 123**,** 739-749.

SAUERBREI, W. & ROYSTON, P. 1999. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 162**,** 71-94.

SCANDINAVIAN STROKE STUDY GROUP 1985. Multicenter trial of hemodilution in ischemic stroke--background and study protocol. *Stroke,* 16**,** 885-90.

SCHAFER, J. L. 1999. Multiple imputation: a primer. *Statistical Methods in Medical Research,* 8**,** 3-15.

SCHIELZETH, H. 2010. Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution,* 1**,** 103-113.

SELLARS, C., BOWIE, L., BAGG, J., SWEENEY, M. P., MILLER, H., TILSTON, J., LANGHORNE, P. & STOTT, D. J. 2007. Risk Factors for Chest Infection in Acute Stroke: A Prospective Cohort Study. *Stroke,* 38**,** 2284-2291.

SENN, S. 2000. The Many Modes of Meta. *Drug Information Journal,* 34**,** 535-549.

SENN, S. 2009. Three things that every medical writer should know about statistics. *Write Stuff,* 18**,** 159-162.

SENN, S., AUCLAIR, P. & JOHNSON, S. 1990. The graphical representation of clinical trials with particular reference to measurements over time. *Statistics in Medicine,* 9**,** 1287-1302.

SENN, S. J. 2007. *Statistical issues in drug development*, Wiley.

SHAH, R. S. & COLE, J. W. 2010. Smoking and stroke: the more you smoke the more you stroke. *Expert Review of Cardiovascular Therapy,* 8**,** 917-932.

SPIEGELHALTER, D., PEARSON, M. & SHORT, I. 2011. Visualizing Uncertainty About the Future. *Science,* 333**,** 1393-1400.

SPIEGELHALTER, D. J. 2008. Understanding Uncertainty. *The Annals of Family Medicine,* 6**,** 196-197.

STAHRENBERG, R., NIEHAUS, C.-F., EDELMANN, F., MENDE, M., WOHLFAHRT, J., WASSER, K., SEEGERS, J., HASENFUß, G., GRÖSCHEL, K. & WACHTER, R. 2013. High-sensitivity troponin assay improves prediction of cardiovascular risk in patients with cerebral ischaemia. *Journal of Neurology, Neurosurgery & Psychiatry*.

STEYERBERG, E. W. 2009. *Clinical Prediction Models: A practical approach to development, validation, and updating*, Springer.

STEYERBERG, E. W., EIJKEMANS, M. J. C., BOERSMA, E. & HABBEMA, J. D. F. 2005. Equally valid models gave divergent predictions for mortality in acute myocardial infarction patients in a comparison of logistic regression models. *Journal of Clinical Epidemiology,* 58**,** 383-390.

STEYERBERG, E. W., EIJKEMANS, M. J. C. & HABBEMA, J. D. F. 1999. Stepwise Selection in Small Data Sets: A Simulation Study of Bias in Logistic Regression Analysis. *Journal of Clinical Epidemiology,* 52**,** 935-942.

STEYERBERG, E. W., EIJKEMANS, M. J. C. & HABBEMA, J. D. F. 2001. Application of Shrinkage Techniques in Logistic Regression Analysis: A Case Study. *Statistica Neerlandica,* 55**,** 76-88.

STEYERBERG, E. W., MOONS, K. G. M., VAN DER WINDT, D. A., HAYDEN, J. A., PEREL, P., SCHROTER, S., RILEY, R. D., HEMINGWAY, H., ALTMAN, D. G. & FOR THE PROGRESS GROUP 2013. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLoS Med,* 10**,** e1001381.

STEYERBERG, E. W., VICKERS, A. J., COOK, N. R., GERDS, T., GONEN, M., OBUCHOWSKI, N., PENCINA, M. J. & KATTAN, M. W. 2010. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology,* 21**,** 128-38.

STONE, C. J. 1986. Comment: Generalized Additive Models. *Statistical Science,* 1**,** 312-314.

STRBIAN, D., ENGELTER, S., MICHEL, P., MERETOJA, A., SEKORANJA, L., AHLHELM, F. J., MUSTANOJA, S., KUZMANOVIC, I., SAIRANEN, T., FORSS, N., CORDIER, M., LYRER, P., KASTE, M. & TATLISUMAK, T. 2012a. Symptomatic intracranial hemorrhage after stroke thrombolysis: The SEDAN Score. *Annals of Neurology,* 71**,** 634-641.

STRBIAN, D., MERETOJA, A., AHLHELM, F. J., PITKÄNIEMI, J., LYRER, P., KASTE, M., ENGELTER, S. & TATLISUMAK, T. 2012b. Predicting outcome of IV thrombolysis–treated ischemic stroke patients: The DRAGON score. *Neurology,* 78**,** 427-432.

STRBIAN, D., MICHEL, P., SEIFFGE, D. J., SAVER, J. L., NUMMINEN, H., MERETOJA, A., MURAO, K., WEDER, B., FORSS, N., PARKKILA, A.-K., ESKANDARI, A., CORDONNIER, C., DAVIS, S. M., ENGELTER, S. T. & TATLISUMAK, T. 2014. Symptomatic Intracranial Hemorrhage After Stroke Thrombolysis: Comparison of Prediction Scores. *Stroke,* 45**,** 752-758.

STRBIAN, D., SEIFFGE, D. J., BREUER, L., NUMMINEN, H., MICHEL, P., MERETOJA, A., COOTE, S., BORDET, R., OBACH, V., WEDER, B.,

JUNG, S., CASO, V., CURTZE, S., OLLIKAINEN, J., LYRER, P. A., ESKANDARI, A., MATTLE, H. P., CHAMORRO, A., LEYS, D., BLADIN, C., DAVIS, S. M., KÖHRMANN, M., ENGELTER, S. T. & TATLISUMAK, T. 2013. Validation of the DRAGON Score in 12 Stroke Centers in Anterior and Posterior Circulation. *Stroke,* 44**,** 2718-2721.

STRIMBU, K. & TAVEL, J. A. 2010. What are biomarkers? *Current opinion in HIV and AIDS,* 5**,** 463.

SUMI, S., ORIGASA, H., HOUKIN, K., TERAYAMA, Y., UCHIYAMA, S., DAIDA, H., SHIGEMATSU, H., GOTO, S., TANAKA, K., MIYAMOTO, S., MINEMATSU, K., MATSUMOTO, M., OKADA, Y., SATO, M. & SUZUKI, N. 2012. A modified Essen stroke risk score for predicting recurrent cardiovascular events: development and validation. *International Journal of Stroke***,** n/a-n/a.

SUN, G.-W., SHOOK, T. L. & KAY, G. L. 1996. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *Journal of Clinical Epidemiology,* 49**,** 907-916.

SUNG, S.-F., CHEN, S. C.-C., LIN, H.-J., CHEN, Y.-W., TSENG, M.-C. & CHEN, C.-H. 2013. Comparison of Risk-scoring Systems in Predicting Symptomatic Intracerebral Hemorrhage After Intravenous Thrombolysis. *Stroke,* 44**,** 1561-1566.

SUZUKI, N., SATO, M., HOUKIN, K., TERAYAMA, Y., UCHIYAMA, S., DAIDA, H., SHIGEMATSU, H., GOTO, S., TANAKA, K., ORIGASA, H., MIYAMOTO, S., MINEMATSU, K., MATSUMOTO, M. & OKADA, Y. 2012. One-Year Atherothrombotic Vascular Events Rates in Outpatients with Recent Non-Cardioembolic Ischemic Stroke: The EVEREST (Effective Vascular Event REduction after STroke) Registry. *Journal of Stroke and Cerebrovascular Diseases,* 21**,** 245-253.

TEALE, E. A., FORSTER, A., MUNYOMBWE, T. & YOUNG, J. B. 2012. A systematic review of case-mix adjustment models for stroke. *Clinical Rehabilitation,* 26**,** 771-786.

THE EC/IC BYPASS STUDY GROUP 1985. The International Cooperative Study of Extracranial/Intracranial Arterial Anastomosis (EC/IC Bypass Study): methodology and entry characteristics. The EC/IC Bypass Study group. *Stroke,* 16**,** 397-406.

THE EUROPEAN STROKE ORGANISATION EXECUTIVE COMMITTEE 2008. Guidelines for Management of Ischaemic Stroke and Transient Ischaemic Attack 2008. *Cerebrovascular Diseases,* 25**,** 457-507.

THE GUSTO INVESTIGATORS 1993. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *N Engl J Med,* 329**,** 673 - 82.

THE IST-3 COLLABORATIVE GROUP 2012a. The benefits and harms of intravenous thrombolysis with recombinant tissue plasminogen activator within 6 h of acute ischaemic stroke (the third international stroke trial [IST-3]): a randomised controlled trial. *The Lancet,* 379**,** 2352-2363.

THE IST-3 COLLABORATIVE GROUP 2012b. Statistical analysis plan for the third International Stroke Trial (IST-3); part of a 'thread' of reports of the trial. *International Journal of Stroke,* 7**,** 186-187.

THE IST-3 COLLABORATIVE GROUP 2013. Effect of thrombolysis with alteplase within 6 h of acute ischaemic stroke on long-term outcomes (the third International Stroke Trial [IST-3]): 18-month follow-up of a randomised controlled trial. *The Lancet Neurology,* 12**,** 768-776.

THE NATIONAL INSTITUTE OF NEUROLOGICAL DISORDERS AND STROKE RT-PA STROKE STUDY GROUP 1995. Tissue Plasminogen Activator for Acute Ischemic Stroke. *New England Journal of Medicine,* 333**,** 1581-1588.

THE STROKE THROMBOLYSIS TRIALISTS' COLLABORATIVE GROUP 2013. Details of a prospective protocol for a collaborative meta-analysis of individual participant data from all randomized trials of intravenous rt-PA vs. control: statistical analysis plan for the Stroke Thrombolysis Trialists' Collaborative meta-analysis. *International Journal of Stroke,* 8**,** 278-283.

THOMPSON, S. G. & HIGGINS, J. P. T. 2005. Can meta-analysis help target interventions at individuals most likely to benefit? *The Lancet,* 365**,** 341-346.

TOLL, D. B., JANSSEN, K. J. M., VERGOUWE, Y. & MOONS, K. G. M. 2008. Validation, updating and impact of clinical prediction rules: A review. *Journal of Clinical Epidemiology,* 61**,** 1085-1094.

TOUZÉ, E., VARENNE, O., CHATELLIER, G., PEYRARD, S., ROTHWELL, P. M. & MAS, J.-L. 2005. Risk of Myocardial Infarction and Vascular Death After Transient Ischemic Attack and Ischemic Stroke: A Systematic Review and Meta-Analysis. *Stroke,* 36**,** 2748-2755.

TREWEEK, S. & ZWARENSTEIN, M. 2009. Making trials matter: pragmatic and explanatory trials and the problem of applicability. *Trials,* 10**,** 37.

TUDUR SMITH, C., DWAN, K., ALTMAN, D. G., CLARKE, M., RILEY, R. & WILLIAMSON, P. R. 2014. Sharing Individual Participant Data from Clinical Trials: An Opinion Survey Regarding the Establishment of a Central Repository. *PLoS ONE,* 9**,** e97886.

VAN BUUREN, S. & GROOTHUIS-OUDSHOORN, K. 2011. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software,* 45**,** 1-67.

VAN CALSTER, B., VAN BELLE, V., VERGOUWE, Y. & STEYERBERG, E. W. 2012. Discrimination ability of prediction models for ordinal outcomes: Relationships between existing measures and a new measure. *Biometrical Journal,* 54**,** 674-685.

VAN DER LEEUW, J., RIDKER, P. M., VAN DER GRAAF, Y. & VISSEREN, F. L. J. 2014. Personalized cardiovascular disease prevention by applying individualized prediction of treatment effects. *European Heart Journal*.

VAN HOUWELINGEN, H. & PUTTER, H. 2012. *Dynamic prediction in clinical survival analysis*, Chapman & Hall/CRC.

VAN HOUWELINGEN, H. C. 2000. Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine,* 19**,** 3401-3415.

VAN HOUWELINGEN, J. C. & LE CESSIE, S. 1990. Predictive value of statistical models. *Statistics in Medicine,* 9**,** 1303-1325.

VAN SWIETEN, J. C., KOUDSTAAL, P. J., VISSER, M. C., SCHOUTEN, H. J. & VAN GIJN, J. 1988. Interobserver agreement for the assessment of handicap in stroke patients. *Stroke,* 19**,** 604-7.

VAN WIJK, I., KAPPELLE, L. J., VAN GIJN, J., KOUDSTAAL, P. J., FRANKE, C. L., VERMEULEN, M., GORTER, J. W. & ALGRA, A. 2005. Long-term survival and vascular event risk after transient ischaemic attack or minor ischaemic stroke: a cohort study. *The Lancet,* 365**,** 2098-2104.

VEERBEEK, J. M., KWAKKEL, G., VAN WEGEN, E. E. H., KET, J. C. F. & HEYMANS, M. W. 2011. Early Prediction of Outcome of Activities of Daily Living After Stroke: A Systematic Review. *Stroke,* 42**,** 1482-1488.

VERGOUWE, Y., ROYSTON, P., MOONS, K. G. M. & ALTMAN, D. G. 2010. Development and validation of a prediction model with missing predictor data: a practical approach. *Journal of Clinical Epidemiology,* 63**,** 205-214.

VICKERS, A. 2006. Whose data set is it anyway? Sharing raw data from randomized trials. *Trials,* 7**,** 15.

VICKERS, A. J. & CRONIN, A. M. 2010. Everything You Always Wanted to Know About Evaluating Prediction Models (But Were Too Afraid to Ask). *Urology,* 76**,** 1298-1301.

VICKERS, A. J. & ELKIN, E. B. 2006. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Medical Decision Making,* 26**,** 565-574.

VICKERS, A. J. & PEPE, M. 2014. Does the Net Reclassification Improvement Help Us Evaluate Models and Markers? *Annals of Internal Medicine,* 160**,** 136-137.

VIECHTBAUER, W. 2010. Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software,* 36**,** 1--48.

VISTA COLLABORATIVE. 2012. *http://www.vista.gla.ac.uk/* [Online]. [Accessed 30/04/2014.

VITTINGHOFF, E. & MCCULLOCH, C. E. 2007. Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression. *American Journal of Epidemiology,* 165**,** 710-718.

VON HIPPEL, P. T. 2007. Regression with missing Ys: An improved strategy for analyzing muiply imputed data. *Sociological Methodology,* 37**,** 83-117.

WAHLGREN, N., AHMED, N., DÁVALOS, A., FORD, G. A., GROND, M., HACKE, W., HENNERICI, M. G., KASTE, M., KUELKENS, S., LARRUE, V., LEES, K. R., ROINE, R. O., SOINNE, L., TONI, D. & VANHOOREN, G. 2007. Thrombolysis with alteplase for acute ischaemic stroke in the Safe Implementation of Thrombolysis in Stroke-Monitoring Study (SITS-MOST): an observational study. *The Lancet,* 369**,** 275-282.

WARDLAW, J. M., MURRAY, V., BERGE, E., DEL ZOPPO, G., SANDERCOCK, P., LINDLEY, R. L. & COHEN, G. 2012a. Recombinant tissue plasminogen activator for acute ischaemic stroke: an updated systematic review and meta-analysis. *The Lancet,* 379**,** 2364-2372.

WARDLAW, J. M., MURRAY, V., BERGE, E. & DEL ZOPPO, G. J. 2013. Thrombolysis for acute ischaemic stroke. *The Cochrane Library*.

WARDLAW, J. M., VON KUMMER, R., CARPENTER, T., PARSONS, M., LINDLEY, R. I., COHEN, G., MURRAY, V., KOBAYASHI, A., PEETERS, A., CHAPPELL, F. & SANDERCOCK, P. A. G. 2012b. Protocol for the perfusion and angiography imaging sub-study of the Third International Stroke Trial (IST-3) of alteplase treatment within six-hours of acute ischemic stroke. *International Journal of Stroke***,** n/a-n/a.

WARLOW, C., SUDLOW, C., DENNIS, M., WARDLAW, J. & SANDERCOCK, P. 2003. Stroke. *The Lancet,* 362**,** 1211-1224.

WEIMAR, C., BENEMANN, J., MICHALSKI, D., MULLER, M., LUCKNER, K., KATSARAVA, Z., WEBER, R., DIENER, H.-C. & GERMAN STROKE STUDY, C. 2010. Prediction of recurrent stroke and vascular death in patients with transient ischemic attack or nondisabling stroke: a prospective comparison of validated prognostic scores. *Stroke,* 41**,** 487-93.

WEIMAR, C., DIENER, H.-C., ALBERTS, M. J., STEG, P. G., BHATT, D. L., WILSON, P. W. F., MAS, J.-L., RÖTHER, J. & INVESTIGATORS, O. B. O. T. R. R. 2009. The Essen Stroke Risk Score Predicts Recurrent Cardiovascular Events. *Stroke,* 40**,** 350-354.

WEIMAR, C., GOERTLER, M., ROTHER, J., RINGELSTEIN, E. B., DARIUS, H., NABAVI, D. G., KIM, I.-H., BENEMANN, J., DIENER, H.-C. & GROUP, S. S. 2008. Predictive value of the Essen Stroke Risk Score and Ankle Brachial Index in acute ischaemic stroke patients from 85 German stroke units. *Journal of Neurology, Neurosurgery & Psychiatry,* 79**,** 1339-43.

WEIMAR, C., KÖNIG, I. R., KRAYWINKEL, K., ZIEGLER, A. & DIENER, H. C. 2004. Age and National Institutes of Health Stroke Scale Score Within 6 Hours After Onset Are Accurate Predictors of Outcome After Cerebral Ischemia: Development and External Validation of Prognostic Models. *Stroke,* 35**,** 158-162.

WEIMAR, C., SIEBLER, M., BRANDT, T., RÖMER, D., ROSIN, L., BRAMLAGE, P. & SANDER, D. 2012. Vascular risk prediction in ischemic stroke patients undergoing in-patient rehabilitation – insights from the investigation of patients with ischemic stroke in neurologic rehabilitation (INSIGHT) registry. *International Journal of Stroke***,** n/a-n/a.

WEIMAR, C., ZIEGLER, A., KÖNIG, I. R., DIENER, H.-C. & COLLABORATORS, O. B. O. T. G. S. S. 2002. Predicting functional outcome and survival after acute ischemic stroke. *Journal of Neurology,* 249**,** 888-895.

WEIR, N., DENNIS, M. S. & ON BEHALF OF THE SCOTTISH STROKE OUTCOMES STUDY GROUP 2001. Towards a National System for Monitoring the Quality of Hospital-Based Stroke Services. *Stroke,* 32**,** 1415-1421.

WEISSCHER, N., VERMEULEN, M., ROOS, Y. B. & HAAN, R. J. 2008. What should be defined as good outcome in stroke trials; a modified Rankin score of 0–1 or 0–2? *Journal of Neurology,* 255**,** 867-874.

WHITE, H. D. & VAN DE WERF, F. J. J. 1998. Thrombolysis for Acute Myocardial Infarction. *Circulation,* 97**,** 1632-1646.

WHITE, I. R. & ROYSTON, P. 2009. Imputing missing covariate values for the Cox model. *Statistics in Medicine,* 28**,** 1982-1998.

WHITELEY, W., JACKSON, C., LEWIS, S., LOWE, G., RUMLEY, A., SANDERCOCK, P., WARDLAW, J., DENNIS, M. & SUDLOW, C. 2009. Inflammatory Markers and Poor Outcome after Stroke: A Prospective Cohort Study and Systematic Review of Interleukin-6. *PLoS Med,* 6**,** e1000145.

WHITELEY, W. N., ADAMS JR, H. P., BATH, P. M. W., BERGE, E., SANDSET, P. M., DENNIS, M., MURRAY, G. D., WONG, K.-S. L. & SANDERCOCK, P. A. G. 2013. Targeted use of heparin, heparinoids, or low-molecular-weight heparin to improve outcome after acute ischaemic stroke: an individual patient data meta-analysis of randomised controlled trials. *The Lancet Neurology,* 12**,** 539-545.

WHITELEY, W. N., SLOT, K. B., FERNANDES, P., SANDERCOCK, P. & WARDLAW, J. 2012. Risk Factors for Intracranial Hemorrhage in Acute Ischemic Stroke Patients Treated With Recombinant Tissue Plasminogen Activator: A Systematic Review and Meta-Analysis of 55 Studies. *Stroke,* 43**,** 2904-2909.

WIJNHOUD, A. D., MAASLAND, L., LINGSMA, H. F., STEYERBERG, E. W., KOUDSTAAL, P. J. & DIPPEL, D. W. J. 2010. Prediction of Major Vascular Events in Patients With Transient Ischemic Attack or Ischemic Stroke. *Stroke,* 41**,** 2178-2185.

WILSON, E. B. 1927. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association,* 22**,** 209-212.

WILSON, P. W. F., D'AGOSTINO, R. B., LEVY, D., BELANGER, A. M., SILBERSHATZ, H. & KANNEL, W. B. 1998. Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation,* 97**,** 1837-1847.

WOLBERS, M., KOLLER, M. T., STEL, V. S., SCHAER, B., JAGER, K. J., LEFFONDRÉ, K. & HEINZE, G. 2014. Competing risks analyses: objectives and approaches. *European Heart Journal*.

WOLBERS, M., KOLLER, M. T., WITTEMAN, J. C. M. & STEYERBERG, E. W. 2009. Prognostic Models With Competing Risks: Methods and Application to Coronary Risk Prediction. *Epidemiology,* 20**,** 555-561 10.1097/EDE.0b013e3181a39056.

YAGHI, S., EISENBERGER, A. & WILLEY, J. Z. 2014. Symptomatic intracerebral hemorrhage in acute ischemic stroke after thrombolysis with intravenous recombinant tissue plasminogen activator: A review of natural history and treatment. *JAMA Neurology,* 71**,** 1181-1185.

ZHANG, N., LIU, G., ZHANG, G., FANG, J., WANG, Y., ZHAO, X., PAN, Y., GUO, L. & WANG, Y. 2013. External Validation of the iScore for Predicting Ischemic Stroke Mortality in Patients in China. *Stroke,* 44**,** 1924-1929.

ZHENG-MING, C. 1997. CAST: randomised placebo-controlled trial of early aspirin use in 20 000 patients with acute ischaemic stroke. *The Lancet,* 349**,** 1641-1649.

ZHOU, X.-H., OBUCHOWSKI, N. A. & MCCLISH, D. K. 2008a. Measures of Diagnostic Accuracy. *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons, Inc.

ZHOU, X. H., LI, C. M. & YANG, Z. 2008b. Improving interval estimation of binomial proportions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences,* 366**,** 2405-2418.